Diplomarbeit

Metagenome Analysis of the Rhizobiome of *Dactylorhiza traunsteineri*

durchgeführt am

Institut für Verfahrenstechnik, Umwelttechnik und Technische Biowissenschaften
der Technischen Universität Wien

betreut von

Univ. Prof. Mag. Dr.rer.nat Robert Mach,
Univ. Ass. Mag.pharm. Gabriel Vignolle und
Univ.Ass. Mag.rer.nat. Dr.rer.nat. Christian Derntl, BSc

durch

Leopold Zehetner, BSc

Wien, 07.09.2021

Author

Leopold Zehetner, BSc

Institution: Vienna University of Technology

Study: Master Technical Chemistry

Specialization: Biotechnology and Bioanalytics

Supervisor

Univ. Prof. Mag. Dr.rer.nat. Robert Mach

Institution: Vienna University of Technology

Department: Technical Chemistry

Institute: Chemical, Environmental and Bioscience Engineering

Co-Supervisor

Univ. Ass. Mag.pharm. Gabriel Vignolle

Institution: Vienna University of Technology

Department: Technical Chemistry

Institute: Chemical, Environmental and Bioscience Engineering

Univ. Ass. Mag.rer.nat. Dr.rer.nat. Christian Derntl, BSc

Institution: Vienna University of Technology

Department: Technical Chemistry

Institute: Chemical, Environmental and Bioscience Engineering

Eidesstaatliche Erklärung

Hiermit erkläre ich eidesstaatlich, dass ich die vorliegende Diplomarbeit eigenständig und ohne fremde Mithilfe verfasst habe.

Wien, 07.09.2021 Leopold Zehetner, BSc

Acknowledgements

Above all, I want to express my deep sense of gratitude to Univ. Prof. Mag. Dr.rer.nat. Robert Mach, head of the Institute of Chemical, Environmental and Bioscience Engineering at the Vienna University of Technology, for his inspirational way of teaching, which encouraged me to perform my master project at his institute and the opportunity to complete my master studies within his institute, despite the Covid-19 pandemic.

Additionally, I would like to sincerely thank my Co-supervisor, Mag. pharm. Gabriel Vignolle, for the inspiring weekly meetings, through which I could gain new knowledge, and his availability, whenever I needed support. His enthusiastic manner encouraged me over the whole time and will have a major impact on my future career. Besides the kind support, which was a cornerstone during this project, I am exceptionally thankful for his advice and the correction and feedback on my thesis.

Furthermore, I am extremely grateful for the motivation and trust, set in me, by Mag.rer.nat. Dr.rer.nat. Christian Derntl, BSc, and Mag. pharm. Gabriel Vignolle to assist in the evaluation of the FunOrder project. To be part of the interesting discussions was a great advantage and a fantastic additional experience.

Besides, I would also like to thank all colleagues and especially the head, Priv. Doz. Dipl. Ing. Dr.techn. Astrid Mach-Aigner, of the group for Synthetic Biology and Molecular Biotechnology within the Institute of Chemical, Environmental and Bioscience Engineering for the great atmosphere and the interesting presentations.

Finally, I want to thank my whole family and friends for their daily support and their belief in my work.

Table of Content

1 Abstract

English Version

Secondary metabolites, such as antibiotics, antioxidants, or other bioactive substances are mainly produced by microorganisms. Genes involved in the production of secondary metabolites in microorganisms are often found clustering together as Biosynthetic gene clusters (BGC), containing the genes for the synthesis of a specific secondary metabolite. To investigate potential BGCs the metagenome of the rhizosphere associated with *Dactylorhiza traunsteineri* was sequenced, assembled, binned and genes were preliminary predicted.

The first aim of my work is to analyze the annotated BGCs containing genes that could be involved in the biosynthesis of unknown bioactive metabolites. Secondly, metagenomic analyses will be performed to predict the potential of the whole population as well as interactions. A third target includes the phylogenetic prediction of the involved organisms for a better understanding of the metagenome. I will use several bioinformatic tools to achieve the aims of this project. Based on the success of my work, future colleagues can express the selected BGCs in model organisms and bioactive secondary metabolites can be investigated.

German Version

Sekundärmetabolite, wie Antibiotika, Antioxidantien oder andere bioaktive Substanzen werden hauptsächlich von Mikroorganismen produziert. Gene, die in deren Produktion involviert sind, werden oft in Biosynthetischen Genclustern (BGC) organisiert vorgefunden. Die Gene eines BGC dienen der Synthese spezifischer Sekundärmetabolite. Um bisher unbekannte BGCs zu entdecken, wurde das Metagenom der Rhizosphere, assoziiert mit *Dactylorhiza traunsteineri,* sequenziert, assembliert, gebinnt und die Gene prognostiziert.

Das erste Ziel meiner Arbeit ist die Analyse und Selektion annotierter BGCs, die unbekannte Gene beinhalten und möglicherweise in die Synthese von bioaktiven Substanzen involviert sind. Als zweites werde ich eine Metagenomanalyse durchführen, um das Potential der Gesamtpopulation und mögliche Interaktionen abschätzen zu können. Das dritte Ziel ist die phylogenetische Analyse der untersuchten Organismen zum besseren Verständnis des Metagenoms.

2 Introduction

2.1 Dactylorhiza traunsteineri

The orchidaceae *D. traunsteineri* (Figure 1) [1] is mainly found in the eastern hemisphere, including marshlands and alpine regions. *D. traunsteineri* is an allotetraploid descendent of the diploid parental strains *D. fuchsia* and *D. incarnata*. This means that *D. traunsteineri* is a hybrid and consists of chromosomes from both parental organisms. [2] Although hybrids are known to be more resistant to environmental changes, due to their larger genomic potential, [3] *D. traunsteineri* is classified as "endangered" by several organizations in central Europe. [4] [5] [6]

Figure 1. The orchidaceae *Dactylorhiza traunsteineri,* [1]

Despite the uncertainty of seasonal effects on rhizobiomes, microbial diversity, and functionality, [7] the impact of climate change on plants including more extreme weather events has been shown. [8] Besides the direct influence on the plants also the environments, i.e. microorganisms in the soil surrounding the plants' roots (rhizobiome), suffer from temperature changes or longer drought or flood periods. [9]

2.2 Rhizosphere

The rhizosphere is known as the space between roots and soil, where microorganisms and plants form a symbiotic relationship (Figure 2). Both, the microorganisms and the plant are dependent on each other and the metabolic exchange. Furthermore, it is known that microbes interact with other microbes within a community in both, symbiotic and pathogenic way. [10] Therefore, a disturbance in the rhizobiome, caused by repeating drought and flooding events due to climate change, could lead to the endangerment of *D. traunsteineri*. [9]

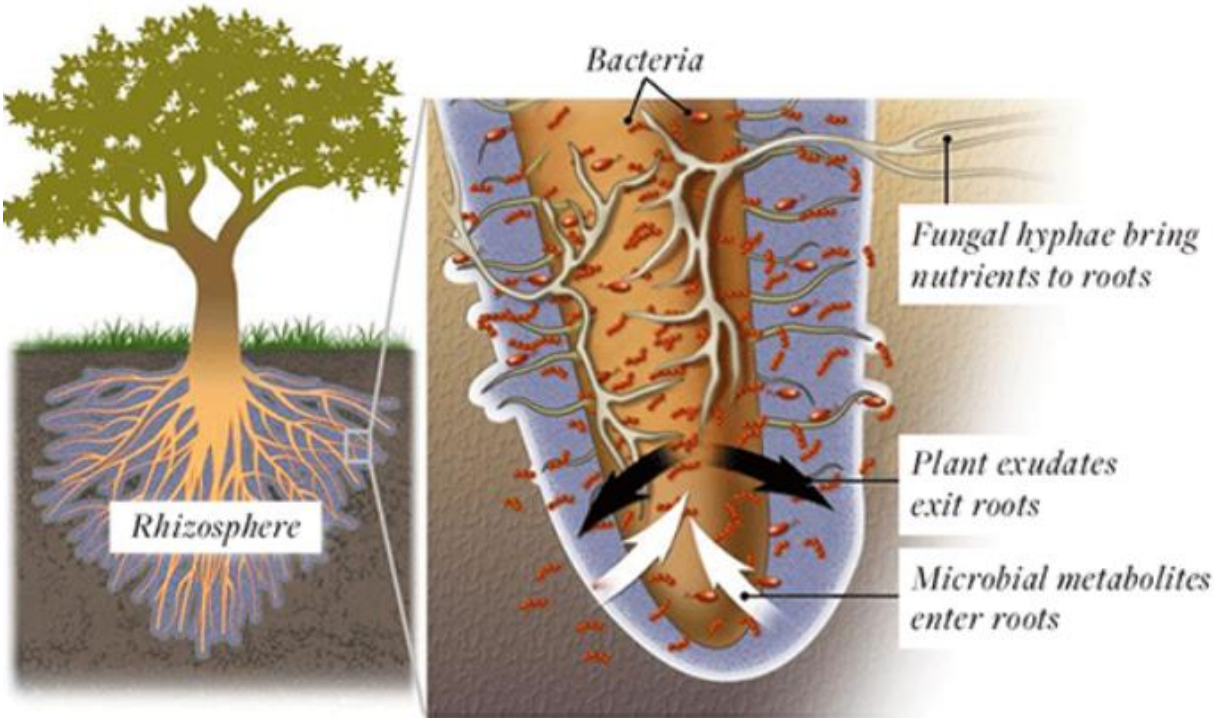


Figure 2. A general overview of the complexity and interaction space between a plant and its rhizobiome, consisting of a diverse group of microorganisms. The symbiosis between plant and rhizobiome is essential for both due to nutrient exchange and protection from potential pathogens. [10]

2.3 Secondary Metabolites

Secondary metabolites (SM) cover a huge and diverse group of compounds, which only have in common that they are not necessary for life-sustaining functions and can therefore be distinguished from the primary metabolism. However, primary metabolites are often used as substrates for the biosynthesis of SM. [11] In contrast to the majority of primary metabolites, SM are often produced under specific circumstances. Therefore, they cover a wide variety of functions, including antimicrobial, pigmenting, antioxidant, or antitumoral, just to name a few examples. [12] Besides their diversity regarding chemistry and biological activity, SM are distributed over all kingdoms, whereby the most SM were detected in plants, fungi, and bacteria [13] Additionally, recent studies investigated, that SM influence host-microbiome interactions, and that these interactions are not limited to exchange of nutrients and primary metabolites. [14] Therefore, more research in this field is needed not just to discover novel bioactive compounds or antibiotics, but also to understand host-microbiome interactions.

2.4 Metagenome and Metagenomics

The complete DNA of a microbial community taken from a certain place or host can be summarized as a metagenome. Until the end of the last century, it was only possible to investigate microorganisms, which could be cultivated. [15] Since the introduction of affordable DNA sequencing methods, it was possible to detect marker genes, i.e. 16S- or 23S-rRNA genes, of uncultivated microorganisms and to expand the tree of life. In contrast to the sequencing of a single cell or organism (genomics), metagenomics is the study of the complete genomic information of a metagenome. [16]

2.4.1 Cultivation-dependent Metagenomics

Culture dependent metagenomics was performed before the introduction of next-generation sequencing technologies. Microorganisms were cultivated under laboratory conditions for investigation under microscopes or to detect the behaviour under changed conditions, i.e. to drugs. [120] Studies after the invention of next-generation sequencing technologies could show, that up to 98% of all microorganisms cannot be cultured under laboratory conditions, due to surrounding factors such as, aerobic/anaerobic conditions, pH ranges, nutrient concentrations, for example. [17] A prominent example for unculturable microorganisms is the group of extremophiles, which are present in all domains of life and populate environments only under certain conditions, for example highly acidic surroundings. [18]

2.4.2 Targeted Metagenomics

To investigate the species diversity efficiently, marker genes, such as 16S-, 18S- or 23S-rDNA genes can be amplified using Polymerase Chain Reaction (PCR), extracted, and sequenced. By this technique the presence (or absence) of certain microorganisms can be detected and, therefore, the composition of the microbiome can be described. Furthermore, it is possible to quantify species within the sample based on the assumption that the selected marker genes are only present once in the genome. [19] A potential problem of targeted sequencing is the generation of chimeras, which are hybrid sequences of at least two parent strands. These hybrids can lead to the incorrect assumption that novel organisms are detected. [20] Therefore, several software tools have been developed to detect and remove chimeric reads from a sample, including ChimericSeq [21] or DADA2. [22]

2.4.3 Shotgun Metagenomics

In contrast to targeted metagenomics, shotgun metagenomics is used to investigate the functional properties of a microbiome. Hence, the complete amount of DNA is extracted from a sample and sequenced. In contrast to targeted sequencing approaches, the sequenced reads must be assembled to contigs and separated by organism using next-generation sequencing platforms and bioinformatic pipelines. Based on shotgun experiments, genomes representations of unculturable organisms can be obtained and analyzed. In consequence, novel proteins, pathways or BGCs can be observed and, in turn, used as a basis for further research, i.e. heterologous expression. [23]

2.5 Biosynthetic Gene Clusters

Biosynthetic gene clusters (BGC) are widely distributed over bacteria, fungi, or plants, containing specialized enzymes for the biosynthesis of bioactive SM. [24] These small compounds cover a variety of functions, including antibacterial, antifungal, antiviral, anti-inflammatory, anticancer, or antioxidant, for example. [25] [26] In a recent study Newman and colleagues investigated that between 1981 and 2014 over 70% of drugs to treat infectious diseases are based on natural products. [27] The enzymes for biosynthesis are often located within a cluster to ensure the concerted expression once the corresponding product is needed. BGCs can be divided into different groups, including non-ribosomal peptide synthetases (NRPS), polyketide synthases (PKS), terpenes, phosphonates, ribosomally synthesized and post-translationally modified peptides (RiPP), or bacteriocins, for instance. The manually curated MIBiG (Minimal Information about Biosynthetic Gene cluster) database contains a total of 1923 BGCs, from which 465 are annotated as complete and 1434 contain at least the minimal information, needed for a BGC, according to MIBiG. [28] [29] The minimal information consists of cluster and compound information, gene information and module information (in case of NRPS and PKS BGCs). [30] During the past decades, especially NRPS and PKS containing BGCs were discovered as a promising source for antimicrobial substances. [25] [31] [32] Both, NRPS and PKS, consist of a various number of consecutive modules which synthesize the final products by the stepwise addition of subunits. Hence, the enzymes can be classified in a starting module, several elongation modules, and a termination module. [33]

In Figure 3. the function of a NRPS is displayed, starting from the corresponding gene, which is expressed to the enzyme, containing three modules, whereby only one elongation module is shown. The biosynthesis is initiated by an adenylation domain (A), which releases pyrophosphate from ATP and binds an amino acid or derivate to the obtained AMP. The substrate is then transferred to the PCP-domain (Peptidyl-Carrier-Domain) by the formation of a thioester bond. A second substrate, incorporated at the neighbouring PCP-domain and also recruited by an adenylation domain, binds to the first by the release of the thioester bonding and, in turn, formation of a peptide bond, catalysed by the condensation domain (C) in between. In the final step, a thioesterase domain binds the precursor through a hydroxy group and releases the molecule after an intramolecular nucleophilic attack. [34] Besides the obligate core modules (A, PCP, C), further modification modules are known, including domains for epimerization, heterocyclization, or methylation. [35] Besides the 21 proteinogenic amino acids, which can be used as substrates, over 500 derivates can also be used as monomeric substrates for non-ribosomal peptide synthesis. [36]

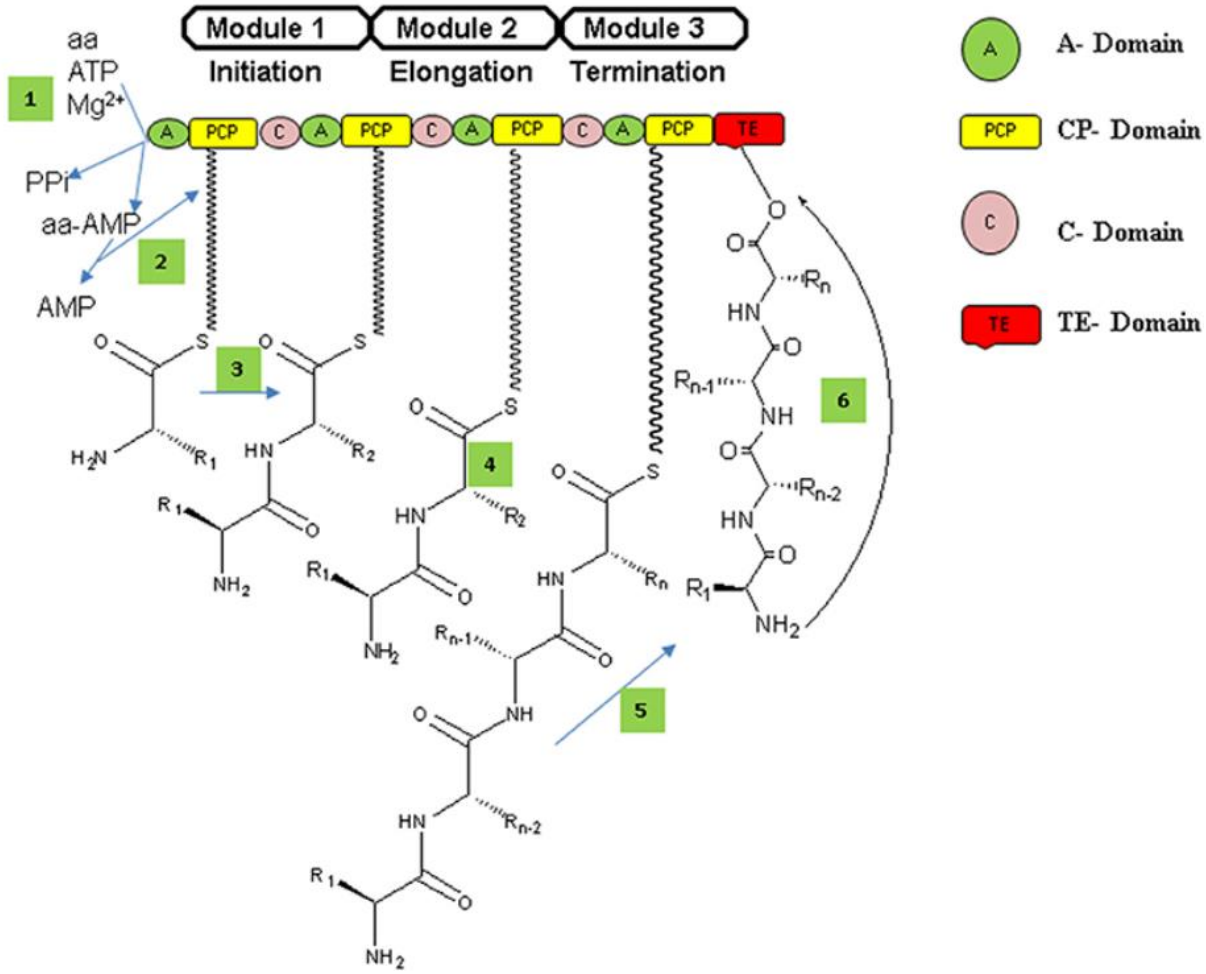


Figure 3. A simplified picture of the general working mechanism of a non-ribosomal peptide synthetase: an amino acid (derivate) is loaded by ATP conversion to AMP. Several elongation modules are necessary to add further substrates. In the final, ring-closing step, the product is released from the thioesterase domain. [34]

In contrast to the wide range of substrates, NRPS can use, PKS only can load certain carbonyl-containing substrates and elongate the polyketide chain through several modules. [37]

In general, PKS starts with an acyltransferase (AT), followed by an acyl-carrier-protein (ACP). After the starting module, several elongation modules, containing an obligate ketosynthase (KS), AT, and ACP domain, are followed by the terminal reductase domain (TD). The release occurs through intramolecular nucleophilic attacks and ring formations, consequently. Besides the core modules, further modification modules can be observed, including dehydratases (DH), ketoreductases (KR) or enoyl reductases (ER). [38] Based on the modules, PKS can be divided into three types, I, II, and III. Type-I PKS consists of noniterative modules, which means that the elongation modules are differently composed, in contrast to type-II and -III enzymes, which comprise the same modules iteratively. The difference between type-II and type-III PKS is the ACP-independency of type-III-PKS. [39]

By combining modules from PKS and NRPS completely novel products can be obtained. In general, two types of hybrid NRPS-PKS products can be distinguished: ones that are generated by a single enzyme by consecutive modules and others, which are synthesized and released by either a PKS or NRPS and are further processed by the other enzyme. Although the underlying chemistry is completely different in these two systems, several bioactive compounds, synthesized by hybrid NRPS-PKS enzymes,[40] could be observed in recent studies, including bleomycin, [41] rapamycin [42] or leinamycin, [43] for instance. Based on the linker hypothesis of modules, it would be possible to interchange PK and NRP modules naturally, which would explain the wide distribution of hybrid NRPS-PKS gene clusters [44]

RiPPs are peptides, which means that after translation certain amino acids are further modified to influence the characteristics of the protein. In prokaryotes, post-translational modifications (PTMs) are relatively rare compared to PTMs in eukaryotes. Therefore, the types and roles of these proteins are of high interest. [45] An important and well-studied subgroup of RiPPs are lantipeptides, which contain specific PTMs and are known for their microzide activity, especially against biofilm formation. They have been observed only in bacteria so far, [46] [47] and were detected in several genera of *Firmicutes, Actinobacteria, Proteobacteria, Bacteriodetes*, and *Cyanobacteria* [48] [49] [50] The specific property all lantipeptides share are post-translationally synthesized sulfide bridges or disulfide bridges. To introduce a sulfide bond, a dehydration step of a hydroxy-group containing, aliphatic amino acid (serine or threonine) is necessary. The resulted 2,3-didehydroalanine (Dha) or (Z)-2,3-didehydrobutyrine (Dhb), respectively, are attacked by the sulfide group of a cysteine residue at the ß-carbon, followed by a protonation step to produce (methyl)lantionine or another conjugation step between dehydroxilized residues, which in turn results in labionin structures. In addition, disulfide bonds between two cysteine residues can be formed. [51]

Lantipeptides can be categorized into four groups based on the biosynthetic pathway. The classes differ due to the number of necessary core enzymes and the biosynthetic pathway. Only class 1 lantipeptides are synthesized by two independent enzymes, LanB and LanC (Figure 4.) [49]

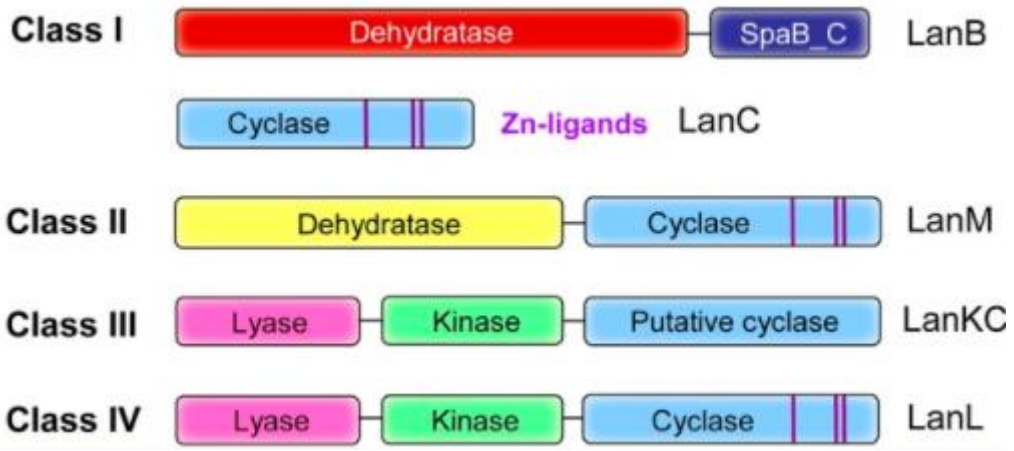


Figure 4. Overview of the four classes of lantipeptide synthesizing enzyme complexes. Only class I consists of more than one enzyme, for the post-translational modifications of the lantipeptides. [49]

Besides the PTMs, all lantipeptides consist of a leader sequence, a core sequence, which is modified, and a follower peptide in some cases. The leader and follower peptides are removed after post-translational modification, most often during the export-process by bifunctional transporter/protease proteins [52] or by separate protease and transporter enzymes.[51]

In addition to characterized bacteriocins, such as maritimacin, and RiPPs, certain BGCs contain uncharacterized core peptides, which are annotated only by the DUF692-domain. DUF-domains (Domain of Unknown Function) are peptides that are common in several proteins but have not been characterized so far. Recent studies characterized members of the DUF692-protein family as potential xylose isomerase [53] or to play a role in hypochlorite detoxification. [54] Besides these findings, van der Donk, et.al., investigated that DUF692-containing proteins could be involved in 3-thiaglutamate synthesis. (Figure 5.) Based on a phylogenetic annotation and clustering of sequences with >40% sequence identity, it was possible to characterize DUF692-proteins as *PmaH*-related. Therefore, DUF692 proteins could be necessary for the ß-carbon excision from the cysteine residue. This rare reaction is followed by two further steps resulting in 3-thiaglutamate. The resulted 3-thiaglutamate was already described as phytotoxin due to its ability to interrupt jasmonate and ethylene signaling pathways. [55]

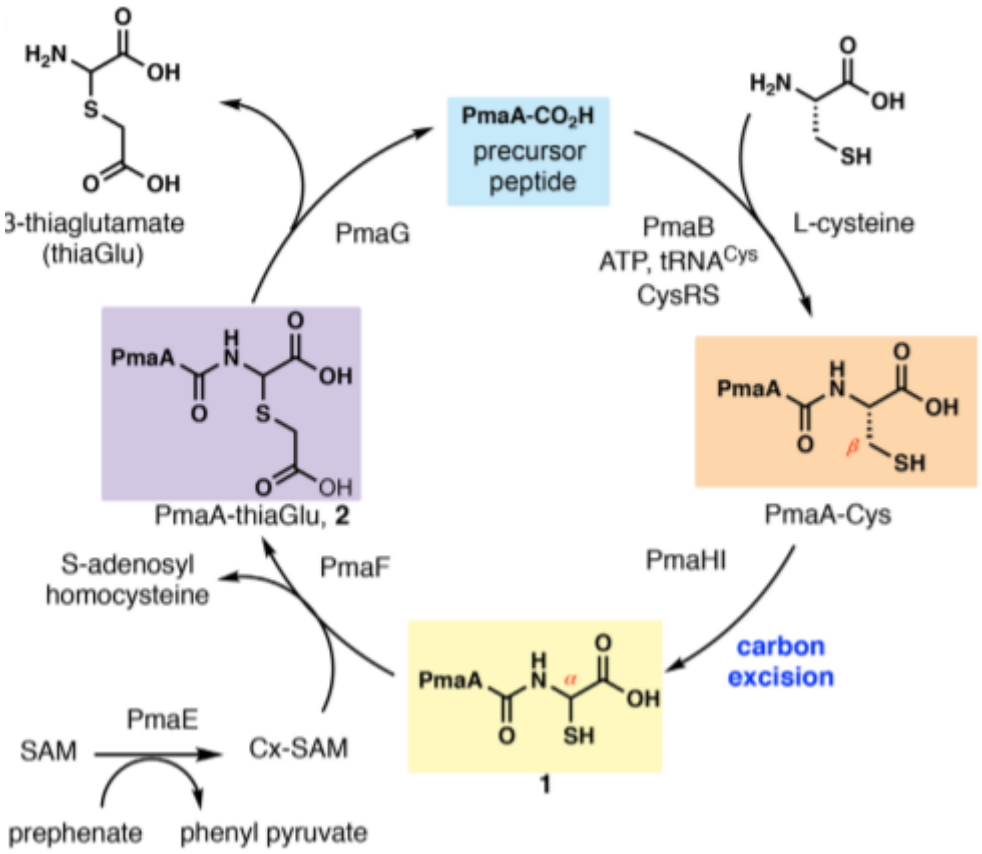


Figure 5. Suggested biosynthesis of 3-thiaglutamate by van der Donk, et.al. The enzyme *PmaH*, involved in the carbon excision step, is thought to contain the DUF692-domain. [55]

Terpenes cover a disparate class of compounds, which are synthesized by the addition of several isoprenoid monomers. The final products are involved in antioxidative reactions or membrane stabilization such as steroids, for instance. Besides the structural variety, they can be modified, i.e. by dehydration, glycosylation, or methylation reactions. [56]

Although the concept of BGCs to produce SM is commonly accepted in the scientific community, a new study suggests that genes, involved in SM-biosynthesis, could also be spread over the genome, but regulated by the same transcription factor. This would increase the SM potential extremely and new strategies, such as weighted coexpression networks, would be needed to investigate differential coexpression. [57]

2.6 Next Generation Sequencing

2.6.1 Sampling and DNA Extraction

The most important step in a sequencing experiment is the first, including sampling and nucleotide extraction. Therefore, several companies developed extraction kits, specific for nucleotide type (DNA, RNA), or the regarding sample source. [58] Based on a shotgun metagenomics approach, it is necessary to include a size selection step after fragmentation of the isolated DNA to obtain only DNA sequences with a certain length. This can be performed after [59] or before adaptor ligation. The size selection is performed using magnetic beads of different sizes. Dependent on the desired DNA length, a size exclusion step above or below the certain size is performed first and discarded, followed by a purification step, covering the remaining DNA fragments. The principle is based on the ability of magnetic beads to be covered by DNA sequences of different lengths, dependent on their concentration. Regarding the SPRIselect system, shorter fragments can be bound with an increasing volume ratio of the beads, whereas longer sequences bind more likely at lower concentrations. [58]

2.6.2 Library Preparation

Before adaptor ligation, the amount of DNA could be increased by amplification using a PCR, especially, when a targeted metagenomic experiment is performed. Adaptor ligation is performed to add terminal DNA fragments on both ends if the sequencing is performed on Illumina platforms. [59]

2.6.3 Sequencing Techniques

For Illumina sequencing platforms, the fragments are hybridized to complementary nucleotide sequences, which are immobilized on a flowcell mediated by the adaptor sequences. The immobilized fragments are amplified through an automated PCR reaction, called bridge building. [59] After bridge amplification is finished, amplified strands bound to one type of adaptor are cut and washed away (for single-end sequencing) and the remaining strand is sequenced by synthesis. In the case of paired-end sequencing, both strands remain hybridized. Sequencing by synthesis means that differentially labelled nucleotides flow over the microarray and are incorporated on the immobilized strand through a PCR reaction by a DNA-polymerase. If a nucleotide is incorporated, the probe is activated through a laser beam and the colour can be detected (Figure 6. [60]). Paired-end read sequencing means that the reads are sequenced from both ends, whereas single-end reading is performed just from one end of the immobilized fragments. Therefore, paired-end read sequencing advances the quality of the sequencing and facilitates the bioinformatic assembly and alignment steps afterward as well as resolving repetitive sequences. [61]

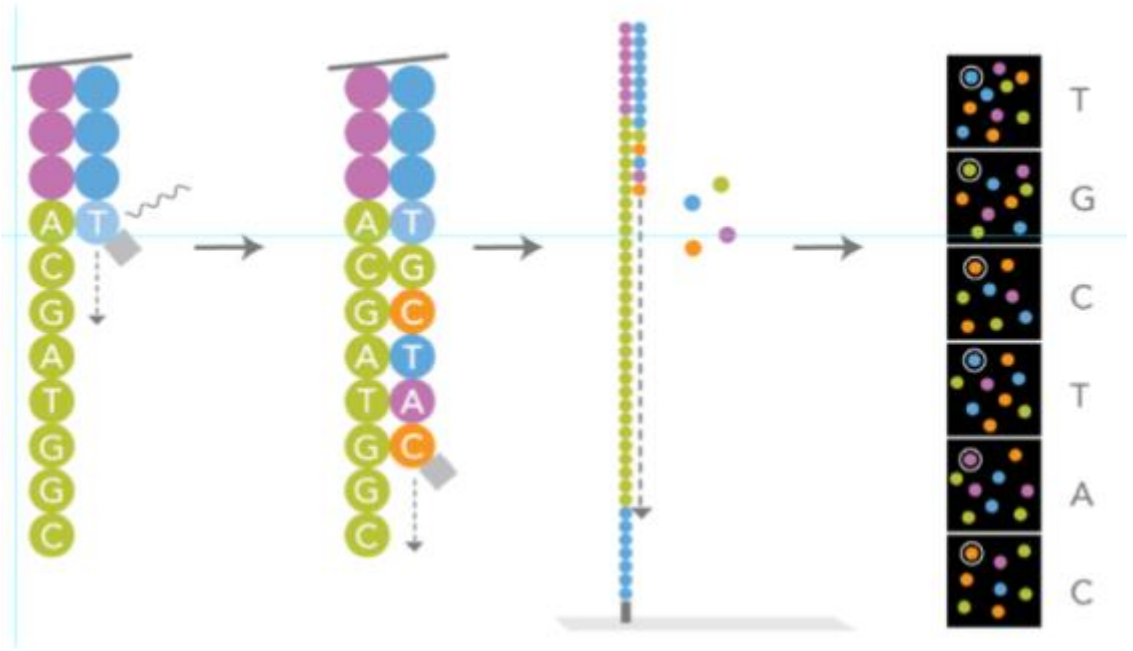


Figure 6. General overview of the sequencing reaction based on an Illumina MiSeq platform. Labeled nucleotides are added through a PCR reaction to the immobilized strand. The nucleotide-specific signal is detected by a laser. [60]

2.6.4 Sequencing of Metagenomes

In contrast to single genome sequencing, DNA is extracted from a sample without prior cell isolation using a metagenomic approach. Consequently, DNA fragments of different cells are mixed and sequenced to obtain metagenome assembled genomes (MAGs) after assembly and binning. This approach does not require prior amplification, which, in turn, leads to low coverage within the reads and therefore, sequencing errors cannot be excluded easily. [62] Recently, it has been investigated that the quality of single genomic approaches and MAG generation does not influence the quality of the genetic functionality of the resulted genomes, which means that metagenomic approaches can be used more efficiently to identify novel organisms. Consequently, MAGs can be used for phylogenetic analyses or investigation of novel enzymes, secondary metabolites or metabolic pathways, for example. [63]

2.7 Metagenome Assembled Genomes (MAGs)

2.7.1 Preprocessing

A metagenome assembled genome (MAG) is defined as a draft genome, resulted from shotgun metagenomics data, followed by computational assembly and binning of the reads. [119] Before the sequenced reads can be used for assembly steps to generate contigs, it is necessary to remove added adaptor sequences at the end of the reads. The software Trimmomatic uses the provided adapted sequences for Illumina sequencing platforms to remove them. Sequence trimming is performed by matching each adaptor sequence to the reads and clip the corresponding sequences off, as far as a sufficient similarity is obtained. Further, Trimmomatic performs a quality trimming of the raw reads. Quality trimming is performed to remove low quality parts of the sequenced reads and, consequently, to increase the quality of the remaining reads. [64]

2.7.2 Assembly

SPAdes is a widely used assembly algorithm, which can be used also for low coverage reads. From the sequenced reads a set of k-mers is generated, which are basic building blocks of different lengths. At the first step, read error correction is performed to discard potential sequencing errors. Therefore, k-mer frequencies are used to resolve bases, which occur seldomly, although low coverage regions cannot be corrected in this way, because all k-mers are present in low amounts and it cannot be distinguished, which bases are erroneous.

After error correction, a de-Bruijn graph is generated, where (k-1) mers represent the nodes and k-mers are edges to connect two nodes, respectively. All equal nodes are merged together to simplify the de-Bruijn graph. A string that consists of all k-mers is called a Eulerian path and would be the final result of the assembly. However, due to not detected sequencing errors and repeating regions in the genome, it is not possible to find “the correct” Eulerian path. Consequently, the graph must be simplified before the correct path can be found. Potential occurring errors include tips, bulges, and chimeras. A tip occurs, when two paths of the de-Bruijn graph fuse in one node and continue from there on as one string. Tips are most often generated by sequencing errors at the terminal side of reads. To remove the tips, SPAdes uses an algorithm for tip clipping in combination with gap closing. When an error is observed in the middle of the reads, a bulge is generated, where two parallel paths develop at one node and merge together again at another node. To overcome this problem, SPAdes uses a bulge removal approach. Finally, a chimeric read consists of two parts of completely different reads and, hence, connects two originally independent paths. This problem is solved by a novel algorithm to discard chimeric reads.

A de-Bruijn graph, including error correction and simplification, is performed iteratively for different k-mer sizes. Using different k-mer sizes implements the advantages of short and long k-mers. Short k-mers lead to more tangled graphs, but resolve errors in low coverage regions, whereas longer k-mers result in more fragmented graphs and are used for repeating and high coverage regions. Repeating regions in genomes are relatively simple to detect, whereas, in terms of nodes, it is difficult to resolve how often a certain repeat occurs originally in the genome. Therefore, a paired de-Bruijn is constructed, based on the paired end reads, which replaces the single (k-1) mers by two (k-1) mers (k-bimers) which are separated by a certain distance. Based on this approach, the paired de Bruijn graph is simplified compared to the original de Bruijn graph. A problem using paired de-Bruijn graphs is that the exact distance between the (k-1) mers of a node is not known. Therefore, the gaps can be close by pairing the reads, from which the (k-1) mer pairs originate. In SPAdes the gap closing step is performed before removing tips or other errors to avoid that wrong tips are clipped only based on coverage.

To address the problem resulted from the unknown distance between the k-bimers of a node, rectangle graphs are introduced, whereby every rectangle is represented by a pair of edges based on an estimated distance in the de-Bruijn graph. By consideration of the insert size, it is possible to investigate the correct path through all rectangles and, hence, assembly the final contig. [65]

The general workflow of the read assembly, using the software MEGAHIT, starts by counting all (k_{min}+1)-mers, resulting in solid and mercy edges, whereby mercy edges are defined as all (k+1)-mers between two solid (k+1)-mers from one read – one without an indegree and one without an outdegree. Mercy edges must be maintained to avoid discarding correct k-mers in low coverage regions, which is especially important in low-depth sequencing of metagenomes. Solid edges are characterized as k-mers which occur 2 times by default. MEGAHIT builds succinct de Bruijn graphs for each k-mer, starting from the smallest to exclude erroneous edges and to fill gaps in low-depth regions up to large k-mers which are necessary for repetitive regions in genomes. After each iteration, bubbles are merged and incorrect edges, as well as edges with low coverage are removed to generate contigs as output. The major advantage of MEGAHIT is its parallel use of CPU (central processing unit) and GPU (graphics processing unit), in contrast to other assembly programs which only use CPUs, which makes MEGAHIT 3-5 fold faster. [66]

2.7.3 Binning

The obtained contigs are separated into kingdoms to remove eukaryotic or host-derived contamination based on their sequence homology. Furthermore, contigs can be assigned to bacterial or archaeal origin. After gene prediction using Prodigal all genes are queried against the non-redundant NCBI database by the accelerated blast implementation Diamond. For each protein, a taxonomy is assigned including the top 10% hits by the majority. The taxonomic classification of the contigs is accepted if a majority of the proteins within the contig can be assigned to a phylogenetic unit, starting from species-level up to phylum. (Figure 7.) Due to horizontal gene transfer over kingdoms, eukaryotic contigs could be characterized as prokaryotic, erroneously. To avoid the incorrectness, differences in coding density can be used to distinguish between eukaryotic (low coding density) and procaryotic (high coding density), whereby it must be considered that incorrect thresholds would remove low-density coding procaryotic genomes. The binning procedure is performed to obtain MAGs. [67]

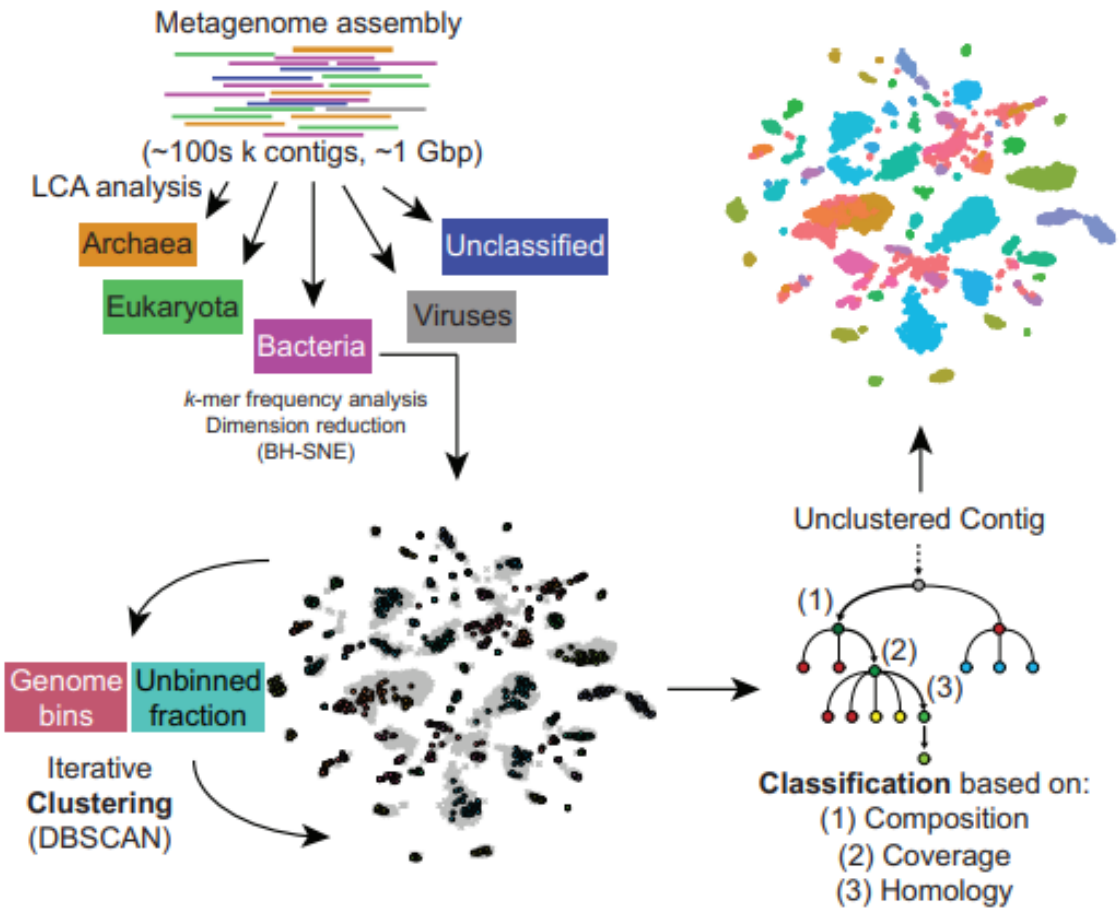


Figure 7. General workflow of the binning software Autometa: the contigs are classified by kingdom and binned through iterative DBSCAN clustering algorithm after the dimension reduction step. The classification is based on composition, coverage, and homology. In contrast to other binning algorithms, Autometa forces not all contigs into bins, which makes the classification more reliable, but results in unclustered contigs as well. [67]

In recent studies, it has been observed that k-mer frequencies show differences between prokaryotic species. [68] [69] Therefore, Autometa performs a principal component analysis (PCA) to obtain a maximum of 50 dimensions based on 5-mer frequencies in contigs after normalization. The dimension reduction is performed using the Barnes-Hut Stochastic Neighbour Embedding (BH-tSNE), followed by the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm for clustering contigs. The DBSCAN algorithm divides data points into core objects (red), directly reachable points within a distance eps from the core points (blue), and not-reachable points (grey). (Figure 8.) [70]) In contrast to alternative clustering methods, such as K-means, it is not necessary to cluster all present points. In addition to the BH-tSNE reduced dimensions, the coverage of the contigs is used as input. The eps parameter defines the radius around each point, starting at 0.3, and is increased by 0.1 iteratively until a final cluster is obtained. The generated clusters are then evaluated and kept according to completeness (>20%) and purity (>90%). The remaining contigs are then used as inputs for another DBSCAN iteration until all contigs are classified or are grouped as unclustered contigs. [72]

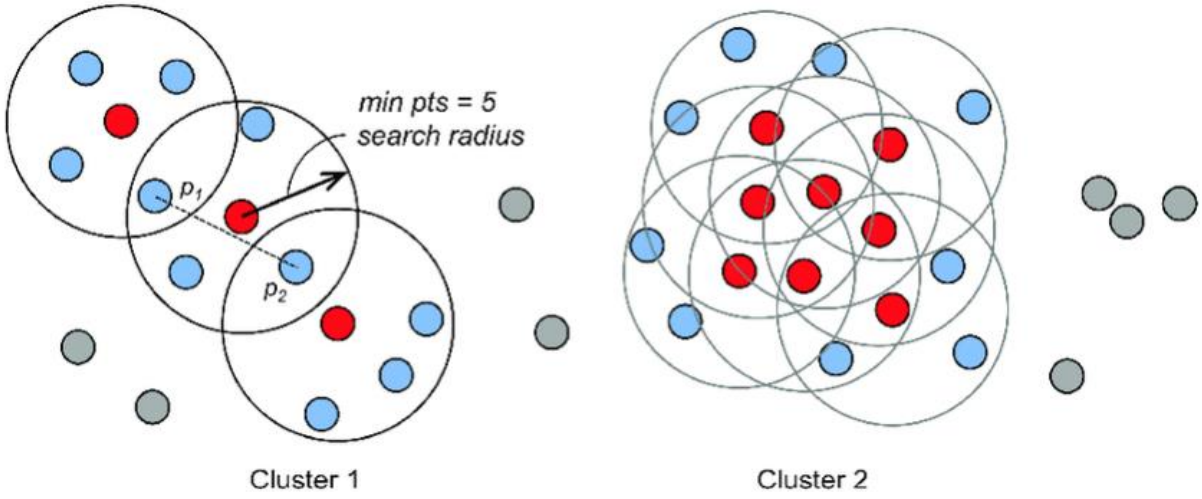


Figure 8. Graphic presentation of the DBSCAN algorithm. Based on core points (contigs) in red, clusters are defined by increasing the eps value after each iteration until no more points can be added. Therefore, some contigs (grey) cannot be assigned to bins and are grouped as unclustered. [70]

To overcome the problem with unclustered contigs by Autometa, a second binning step is needed. MetaBAT2 is based on a more sensitive clustering algorithm. It computes normalized 4-mer frequency (TNF) scores. The TNF scores are normalized by quantiles using the abundance score to obtain a composite score S:

$$S = sqrt(TNF^{1-w} * ABD^{w} * COR)$$

where ABD is the abundance score, COR is an ABD correlation score using Pearson correlation and w is calculated through:

$$w = \frac{nABD}{nABD + 1}$$

with nABD as the number of samples with sufficient coverage, which is set to >1 by default. Due to the inaccurate separation of related species by TNF, S can be used to differentiate between them. The graph-based clustering algorithm assigns each contig to a node and edges are based on the similarity of the nodes. The initial graph is generated only based on TNF scores including a limited number of edges per node to avoid too long running times, followed by iterative graph building cycles, whereby edges with the highest S value are used for the following graph partition. [72]

This is performed by a modified label propagation algorithm (LPA), [71] which is used to identify communities in network structures. Therefore, each node is labeled and the network is reordered based on the S score for each edge. For each node, the most frequent label in the neighbourhood is returned. Consequently, clusters that are interconnected with high density obtain equal labeling earlier. Finally, Fisher's method is performed to determine to which neighbourhood the contig belongs most probably. [72]

2.7.4 Evaluation of MAGs

To evaluate the quality of novel sequenced organisms the software CheckM can be used. CheckM uses universal marker genes which are detected only once in >97% of the lineage-dependent reference genomes. The collocation of two marker genes is determined by the occurrence of both within 5 kbp in >95% of the genomes of one lineage. Based on the sets of collocated marker genes the completeness is calculated as the ratio of detected marker gene sets divided by all marker gene sets. The genome contamination is determined by the number of multiple copies of marker genes within each set, divided by all sets of marker genes. Finally, a heterogeneity score is calculated as the ratio between gene pairs, which occur multiple times and exceed an amino acid identity of 0.9. [73]

QUAST (quality assessment tool for genome assemblies) is used to obtain characteristic values of each MAG, including N50, L50, number of contigs, total length, GC-content, or average coverage. Though the program can compute much more metrices, here are only those described, which were used for the quality assessment: [74]

N50 value is the length of the shortest contig of a set of longest contigs that represent 50% of the genome.

L50 value describes the number of contigs, beginning from the longest, to cover 50% of the genome.

The GC content is calculated by the number of guanines and cytosines, divided by the total length of the genome in bases.

The average coverage is the mean of coverages from all contigs and is displayed as a histogram by QUAST.

2.8 Phylogeny and Taxonomy

2.8.1 Last Common Ancestor

It has been established that all life forms can be divided into three domains of life, including *Archaea*, *Bacteria*, and *Eukarya*. [87] Based on the tree of life, related organisms can be condensed into clades, as a part of the phylogenetic tree with a common ancestor. Such a group of organisms is defined as monophyletic. [88] The last common ancestor (LCA), or most recent common ancestor (MRCA), of a clade is thought to contain shared properties of organisms within the clade. [89] Therefore, it is possible to characterize novel organisms based on their contigs in comparison to references, as described earlier, using the Autometa program. [67]

2.8.2 Ribosomal Protein Analysis

Instead of the most commonly used phylogenetic analysis based on 16S- and 23S-rRNA genes, it has been shown in several studies, [90] that over 50 ribosomal proteins can be identified, from which 34 are universally conserved and 23 ribosomal proteins are specific for bacteria. [91] In contrast to the rRNA approaches, in the phylogenetic analysis based on ribosomal proteins potential chimeric artifacts are avoided. [90] For this purpose, ribosomal protein subunits of the same type must be present in all organisms. A potential disadvantage is that some ribosomal proteins consist of relatively short amino acid sequences, which makes it difficult to predict correctly and, hence, it is possible that they are not annotated. The ribosomal proteins are then used to calculate the phylogenetic relation. [91]

2.8.3 Average Nucleotide Identity

The Average Nucleotide Identity (ANI) is a robust method to calculate nucleotide sequence similarities of whole genomes between two or more species. Organisms from the same species share an ANI of >95% and showing an average nucleotide identity of >83% pertain inter-species related genomes. Since the first *in silico* approach of Goris and colleagues using the BLASTN program, [92] several other algorithms have been established for calculating the average nucleotide identity based on MUMmer (ANIm) or USEARCH (OrthiANIb). [93] [94] [95] Due to the increasing amount of genomes, the most important step is to efficiently calculate pairwise ANI values for a huge number of organisms while maintaining accuracy. In contrast to ANIm and OrthoANIb, which were investigated to be significantly faster than the BLASTN algorithm for ANI scores above 90%, [100] FastANI reduces the computation time 2-3 fold and also covers genomes with lower ANI scores. At first, fragments of size l of the query genome (A) are generated avoiding overlaps, followed by a mapping step of the fragments to a reference genome (B). Mashmap uses a winnowed-MinHash estimator for alignment prediction instead of performing the alignment and results in triplets, containing the fragment index, the identity estimation value, and the starting position in the reference genome. From each fragment, only the triplet with the highest estimated identity is saved. Based on these triplets, a reciprocal triplet set of the reference genome is generated, containing only the highest estimated identity for each position in the reference genome. The identity values from the reciprocal set are used for the mean calculation to obtain the final ANI score. (Figure 9.) FastANI uses a fragment length of 3 kbp, a minimal amount of 50 reciprocal mappings, and an identity cutoff of 80% by default to ensure the low runtime and high estimation accuracy. Instead of using direct alignments, the underlying Mashmap algorithms generate a relation for the alignment identity of two sequences and the Jaccard similarity based on a Poisson model of the single k-mers. [97]

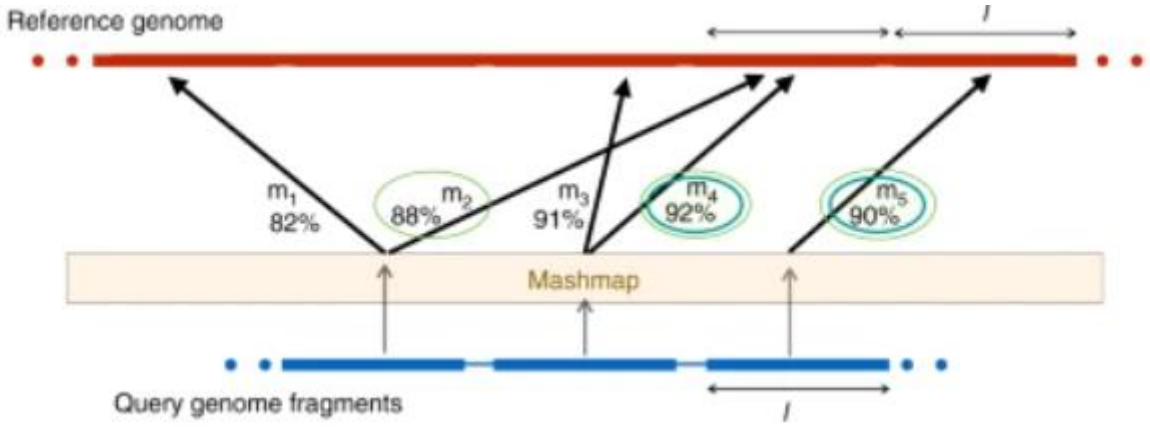


Figure 9. Simplified presentation of the Mashmap algorithm, implemented in the FastANI software. [97]

2.8.4 Phylogenetic Tree

Phylogenetic trees are used to display the evolutionary relationship of organisms. To create a phylogenetic tree, several approaches can be used, including 16S- and 23S-rRNA analysis,[90] ribosomal protein comparison [91] or based on the whole proteome of organisms. The latter strategy is the most challenging due to the comparison of whole proteomes instead of certain selected proteins or DNA sequences. OrthoFinder is a program to address this problem and to generate phylogenetic trees with high reliability. At first, proteins are separated into sets of orthologs (orthogroups). Orthologs are genes, which were already present in the last common ancestor. Based on each orthogroup an unrooted tree is created using DendroBLAST, followed by generating an unrooted tree for the species using the STAG algorithm. The STAG (species tree from all genes) program uses not only one-to-one orthologs but most closely related genes of orthogroups. The unrooted tree on the level of the organism is then rooted using the STRIDE (Species Tree Root Inference from Duplication Events) algorithm to characterize gene-duplications in the orthogroup-based trees. Based on the rooted species tree, the unrooted orthogroup trees are rooted, in turn. Furthermore, gene duplication and loss events are detected by a duplication-loss-coalescence model and the species-overlap method. A duplicated gene leads to the occurrence of two or more genes within an orthogroup and, in turn, results in a false negative (if the true ortholog is not detected) or false positives (if the paralog is identified as an ortholog). [98] Paralogs are generated after gene duplication events, whereas orthologs are generated through speciation events. [106]

2.9 Postprocessing Analysis

2.9.1 Gene and Protein Prediction

Prodigal (Prokaryotic Gene Recognition and Translation Initiation Site Identification) is a software tool for gene and peptide prediction in prokaryotes. At first all start and stop codons in the genome are detected, including only standard codons, such as ATG, GTG, or TTG as starting codons, for instance. Based on the start codons, all open reading frames (ORFs) are predicted, and a frame bias model is used to score the start codons, based on the length of ORFs and the G/C occurrence at each codon position. An ORF is defined as a region between start and stop codons, which can potentially be transcribed to mRNA and translated to a protein. Before the first dynamic programming step, the starting codons (nodes) are selected, which comply with the highest-scoring values and overlap a stop codon within fewer than 60bp.

The coding score is calculated for each 6-mer by the logarithmic fraction of percentage occurrence in the training set and the percentage occurrence within the sequence. The final coding score is equal to the sum of all coding scores of the 6-mers within a gene. The coding score represents the likelihood of sequences between a start and a stop codon to be a gene. Therefore, it is used for determination of potential genes.

During the dynamic programming step, genes are selected, when nodes, assigned to starting codons, reach stop codons and intergenic sequences are predicted vice versa. Next, all possible 6-mers of all sequences are generated and used to calculate coding scores, after log transformation. For the predicted genes ribosomal binding sites and shine Dalgarno motifs are predicted over 10 iterations. Consequently, a final score is calculated based on the start and coding score. If a gene length is below 250 bp it is penalized. Finally, a second dynamic programming step is made based on 6-mers to eliminate negative scores. This leads to a dataset, containing gene coordinates and translated proteins. [75]

2.9.2 Metabolic Pathway Analysis

2.9.2.1 Quorum Sensing

Quorum sensing (QS) is used by microorganisms to communicate with each other through small molecules. These molecules are involved in gene regulation and therefore, phenomena including, bioluminescence or biofilm formation can be controlled, based on cell population density. [76] Several main signalling molecules have been studied in the past, including different types of autoinducer (AI-1, AH-2, and AI-3) [77] [78] [79] or autoinducerpeptides. [80] The signal can activate the transcription of specific proteins, including BGCs to react to the population density. [76] [81]

2.9.2.2 Alkansulfon Metabolism

Sulfonates are a special class of molecules, containing a C-S bonding. These bonds must be synthesized by specific enzymes. Two well-studied representatives are allicin, the antimicrobial compound produced in garlic [82] and coenzyme M and B, which are involved in methane metabolism, most commonly in archaea. [83] Although, coenzyme M is a central molecule in methane metabolism the biosynthetic pathway is not completely understood. Recently, it has been investigated that there exist alternative biosynthetic pathways in different phyla, including *Alphaproteobacteria,* for instance. Independent from species, the final step in coenzyme M biosynthesis is still unknown, although it is assumed that an enzyme, containing a 4Fe-4S cluster, is involved. [84]

2.9.3 Prediction of Biosynthetic Gene Clusters

antiSMASH is a data mining software, which is used to predict biosynthetic gene clusters and is based on searches against profile hidden Markov models (pHMM). These were generated by multiple sequence alignments of already described core proteins or protein domains of BGCs from databases, including the antiSMASH database and the MIBiG (Minimal information about biosynthetic gene clusters) repository. Together with novel pHMMs based on seed alignments (BLAST), a library is obtained and used to predict query proteins. The final BGCs are then generated by the inclusion of all genes within 5, 10, or 20 kb from the core enzyme. Based on this greedy approach it is possible that 'superclusters' are predicted if two core enzymes of different BGC are located within the inclusion distance. By this approach, only known BGCs can be predicted. To also identify novel BGCs, the input sequences are separated into protein families by the Pfam program to obtain Pfam domains, which in turn are used as input for an HMM to distinguish between 'gene cluster'- and 'rest-of-the-genome'-states. In the used version, antiSMASH 4, new algorithms for module and product prediction for NRPS-, PKS-, RiPP- and terpene-BGCs are integrated. [85] [86]

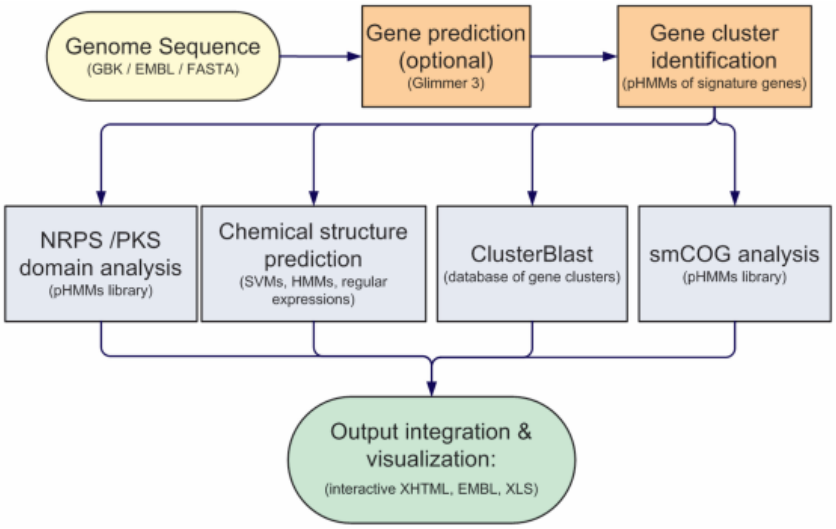


Figure 10. General Workflow of the antiSMASH software for prediction of Biosynthetic Gene Clusters. [85]

2.10 Aims and Expectations

In this work, our aim is to create the first shotgun metagenome of the rhizobiome from the endangered orchid *D. traunsteineri*. Therefore, phylogenetic and functional analyses were performed to obtain information about the microbiome, as well as the biologic potential. The analysis of the generated MAGs was performed with a special focus on potential bioactive SM produced by BGCs. This sets the stage for future analysis of the holobiont. The most promising BGCs might be further used as candidates for heterologous expression.

3 Materials and Methods

3.1 Sequencing

The whole sampling, sample preparation and sequencing procedure was performed following in house protocols.

3.2 Metagenome Assembly

3.2.1 Preprocessing

The sequenced reads were obtained in the forward and reverse directions. Terminal barcode sequences were removed by the software Trimmomatic (v0.40). [64]

3.2.2 Assembly and Binning

The trimmed reads from both samples were split into three datasets: one containing reads from the soil, one containing reads from the washing water, and one containing the remaining reads from both samples. All three datasets were assembled with the software MEGAHIT (v1.2.9) [66] and the generated contigs from all three datasets were combined. To generate metagenome-assembled genomes contigs were binned using the software Autometa (v2.0) [67] and MetaBAT2 (v1.7) [72] for unclustered contigs. For further improvement, the corresponding reads of each MAGs' contigs were extracted and reassembled with SPAdes (v3.14.1), [65] using the meta option.

3.2.3 Evaluation of MAGs

The quality of the MAGs was evaluated by the software CheckM (v1.1.3) [73] and QUAST (v5.0.2). [74] Based on the results from CheckM [73] it was possible to classify all MAGs by completeness and contamination. The MAGs were divided into three groups by certain threshold values. The classification rules are displayed in Table 1. Furthermore, the heterogeneity for each MAG was obtained.

Table 1.: Classification thresholds based on completeness and contamination levels, obtained from CheckM [73] to group MAGs.

CLASSIFICATION OF MAGS	COMPLETENESS [%]	CONTAMINATION [%]
HIGH QUALITY	> 75	< 25
MEDIUM QUALITY	30-75	< 25
LOW QUALITY	< 30	> 25

The QUAST [74] analysis was also performed for each MAG and general statistics, including N50, L50, contig length, average coverage, and GC content, were generated to obtain an assessment of the sequencing and binning quality. Based on these results, the MAGs were further characterized.

3.3 Phylogeny and Taxonomy

For the taxonomic analysis of the generated MAGs, several approaches were used: at first it was tried to identify equally annotated ribosomal proteins to set them into relation. Furthermore, a taxonomic table was created by the software Autometa [67] , based on the non-redundant database (NCBI). As a further attempt, the average nucleotide identity was calculated using the software FastANI (v1.33) [97] and, finally, phylogenetic trees were generated based on the software OrthoFinder (v2.5.2). [98]

3.3.1 Ribosomal Protein Analysis

All ribosomal proteins were extracted for each MAG based on the KEGG [107] and PANNZER2 [108] annotation. The obtained proteins were quantified by type to identify the most abundant representative in all MAGs.

3.3.2 Lowest Common Ancestor

The lowest common ancestor was predicted using the corresponding option from Autometa. The binned MAGs were used as input and for each MAG a taxonomic prediction was obtained, except those MAGs, which could not be clustered by Autometa and were grouped as unclustered MAGs. [67]

3.3.3 Average Nucleotide Identity

A generated library of 199 reference organisms, downloaded from the NCBI-database was generated and used to calculate the average nucleotide identity of the MAGs using the software FastANI. [97] The reference genomes were classified by phylum and an overview of the distribution of reference genomes is displayed in Figure 11. For each MAG the highest ANI-score was extracted. Furthermore, the relative abundance of all ANI-scores for each MAG was calculated.

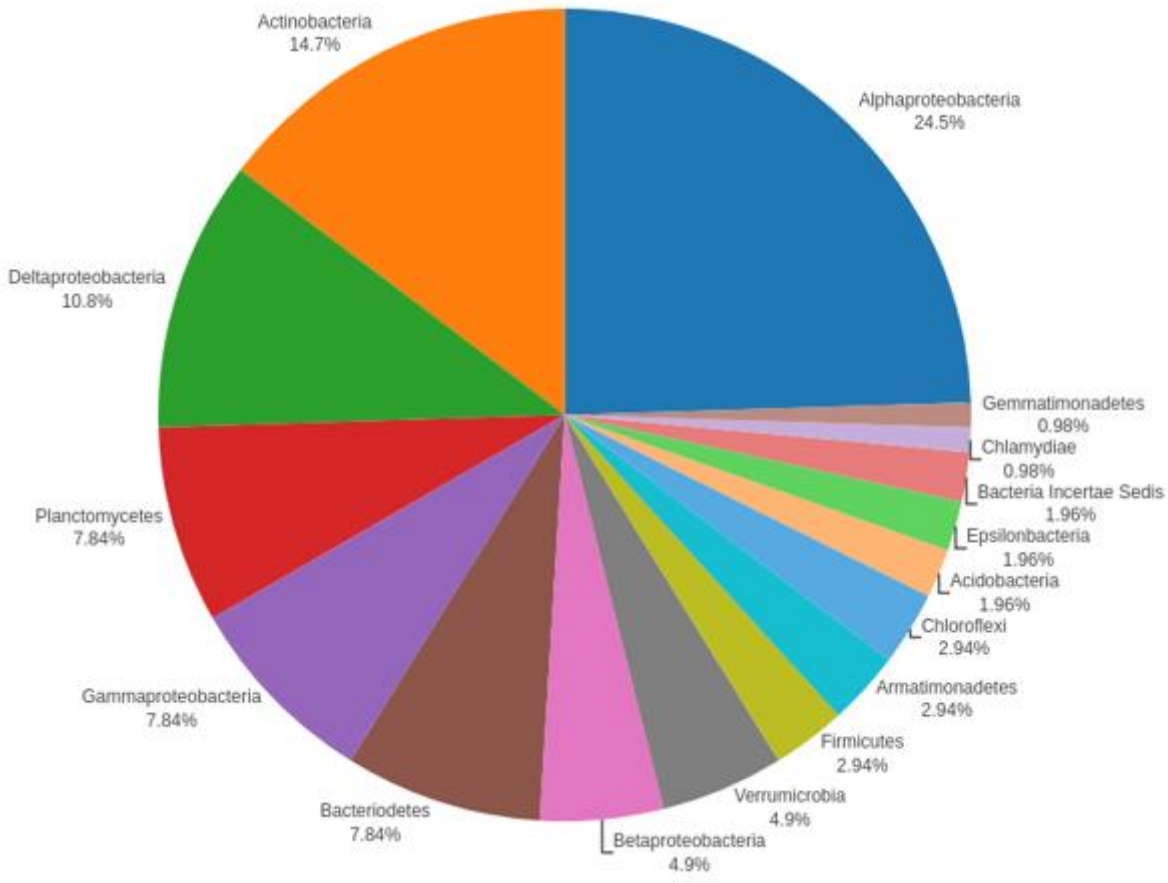


Figure 11. Abundance of the reference genomes used for average nucleotide identity analysis categorized by phylum.

3.3.4 Phylogenetic Tree

A phylogenetic tree was constructed based on the results of the program OrthoFinder. The number of reference genomes was reduced to 21 due to computational limitations. The selected organisms were combined with the MAGs with high and medium quality before running OrthoFinder. The resulted file in Newick Format contained node values as well as distance-dependent edge lengths. To display the phylogenetic tree graphically, the online tool phylo.io [98] was used for rooted trees and the R-package ggtree (v3.0.4) [100] was used to create an unrooted tree. Furthermore, it was tried to classify the MAGs more specifically by adding other reference genomes.

3.4 Postprocessing Analysis

3.4.1 Gene Prediction and Annotation

For each generated MAG as well as for all generated contigs from the rhizobiome, genes and proteins were predicted using the software Prodigal (v2.6.3). [75] The fasta files containing the peptide sequences were annotated by the KEGG [107] (E-value = 0.01) and PANNZER2 [108] databases. The KEGG [107] annotation was further used to investigate metabolic pathways, whereas PANNZER2 [108] and the corresponding domains and gene ontology were annotated additionally. The final functional annotation was based on both database annotations, but with stronger consideration of the KEGG [107] annotation. For further evaluation of unspecific annotations, domain prediction and gene ontologies were taken into account to provide a reliable description. For the rhizobiome, the corresponding protein files were split into several files for annotation due to the limited number of entries from the databases.

3.4.2 Metabolic Pathway Analysis

For pathway analyses, the KEGG database [107] was used to investigate the metabolic potential of the MAGs with high or medium quality. For missing enzymes in a pathway, it was tried to find corresponding proteins based on the PANNZER2 [108] annotation. Besides primary metabolic pathways, some MAG-specific pathways were described in more detail, including enediyne biosynthesis, QS, and alkanesulfone biosynthesis.

3.4.3 Prediction of Biosynthetic Gene Clusters

For each MAG and the rhizobiome biosynthetic gene clusters were predicted using the software antiSMASH (v4.2.0). [85] The BGCs were investigated for MAGs from all qualities and an overview by BGC type and each MAG was generated. As a next step, BGCs from specific types, including NRPS, PKS, terpenes, bacteriocins, or lantipeptides, were annotated based on the KEGG [107] and PANNZER2 [108] annotations and checked for completeness. Besides the involved genes, module prediction of the NRPS and PKS clusters were investigated for their completeness and divided into three groups: complete modules, at least one complete module, and no complete modules. Based on these results 10 BGCs were selected and described for future investigation.

4 Results

4.1 Metagenome Assembled Genomes

4.1.1 Generation of MAGs

The binning process using Autometa [67] resulted in 37 predicted MAGs and a group of contigs classified as “unclustered”. These contigs were then rebinned with MetaBAT2 [72] and resulted in ten additional MAGs. From 47 generated MAGs, it was possible to reassemble 32 MAGs, whereas 15 could not be reassembled using SPAdes. [65] Overall, 1,999,537 contigs were generated after the assembly using MEGAHIT [66] , from which 21,943 could be binned to MAGs.

4.1.2 Quality Assessment

The contamination (Figure 12.) and heterogeneity (Figure 13.) of each MAG are plotted against the completeness to visualize the classification of the MAGs by quality. From the defined thresholds in Table 1., five MAGs with high quality, eight MAGs with medium quality, and 27 MAGs with low quality were obtained. Seven of the originally generated MAGs were discarded after quality evaluation, due to a contamination score of > 100% or completeness below 10%.

In Table 6. an overview of the final classification of the MAGs, sorted by completeness is displayed. Furthermore, values for contamination and heterogeneity are also displayed.

Figure 12. Graphic presentation of MAGs. Based on their completeness (x-axis) and contamination (y-axis), the quality classification was assigned.

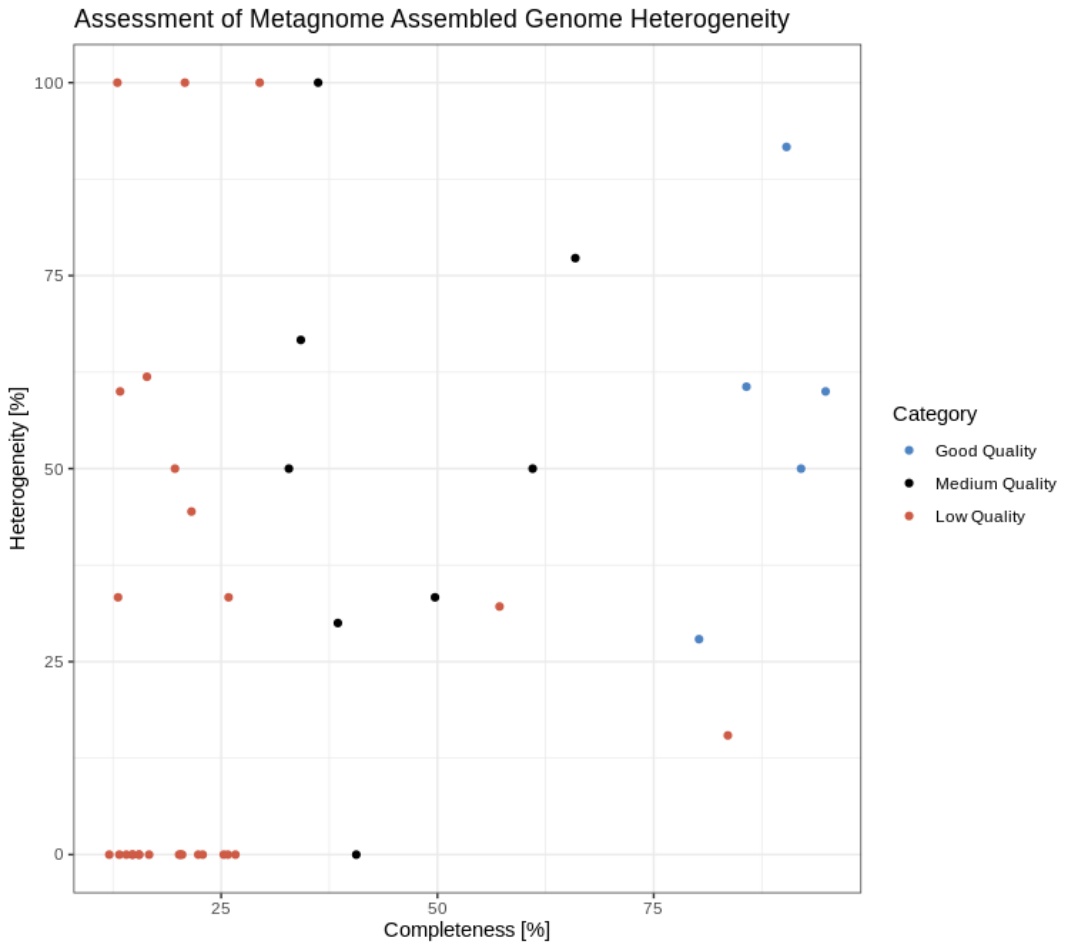


Figure 13. Graphic classification of MAGs. Based on their completeness (x-axis) and heterogeneity (y-axis), an overview of the potential novelty of the MAGs is obtained.

Besides completeness and contamination scores, the heterogeneity allowed a first impression of the novelty of the MAGs, though, it had to be considered that high heterogeneity values could also result from two or more organisms that were grouped as one MAG during the binning process.

Based on the quality assessment MAGs were further characterized according to their N50, L50, average coverage, number of contigs, and GC content. The N50 (Figure 14.) and L50 (Figure 34.) values as well as contig amounts (Figure 15.) of the MAGs confirmed the results from the quality assessment. Generally, it was observed that MAGs with high quality were assembled by longer contigs, indicated by higher N50 and lower L50 values, and a lower amount of contigs to cover the genomes, consequently. In MAG_Dt_26, the best assembly statistics could be found with just 49 contigs and the highest N50 value of 221,190 bases. In addition, high-quality MAGs were identified to have a higher average coverage overall contigs compared to low or medium quality MAGs (Figure 16.). Especially, MAG_Dt_25 and MAG_Dt_26 were observed to comprise the highest average coverage values. Finally, average GC contents were displayed in Figure 35. for each MAG.

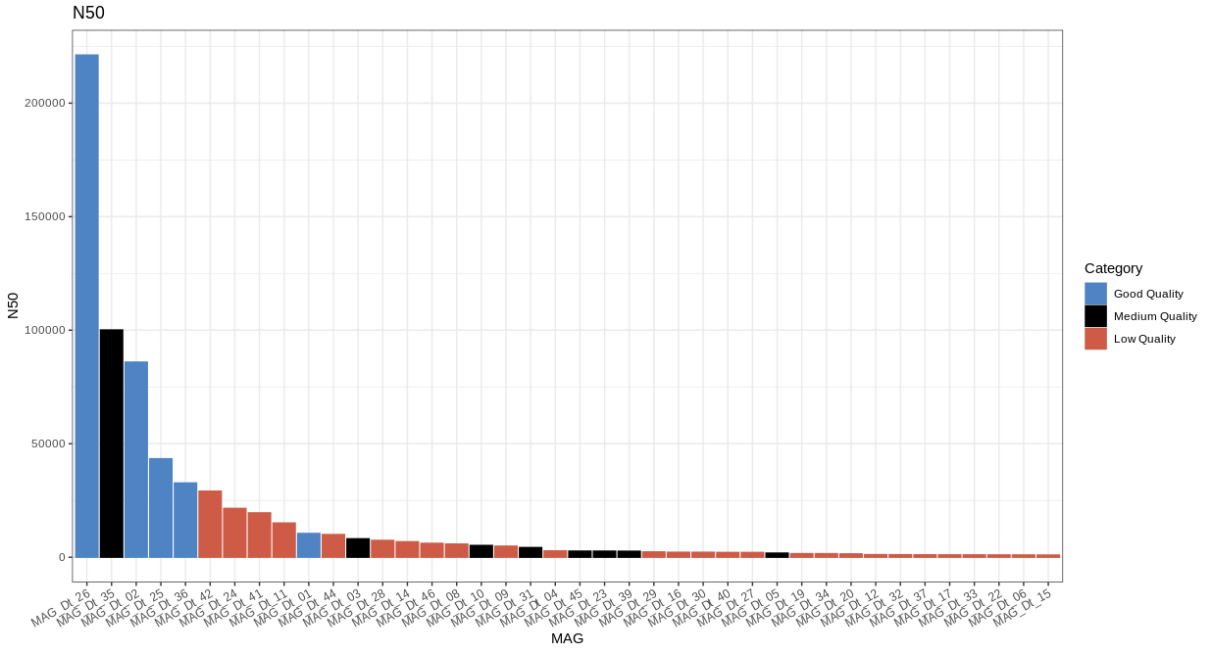


Figure 14. N50 values, sorted by length in bp and assigned to each MAG. High- and medium quality MAGs show generally higher N50 values than low-quality MAGs.

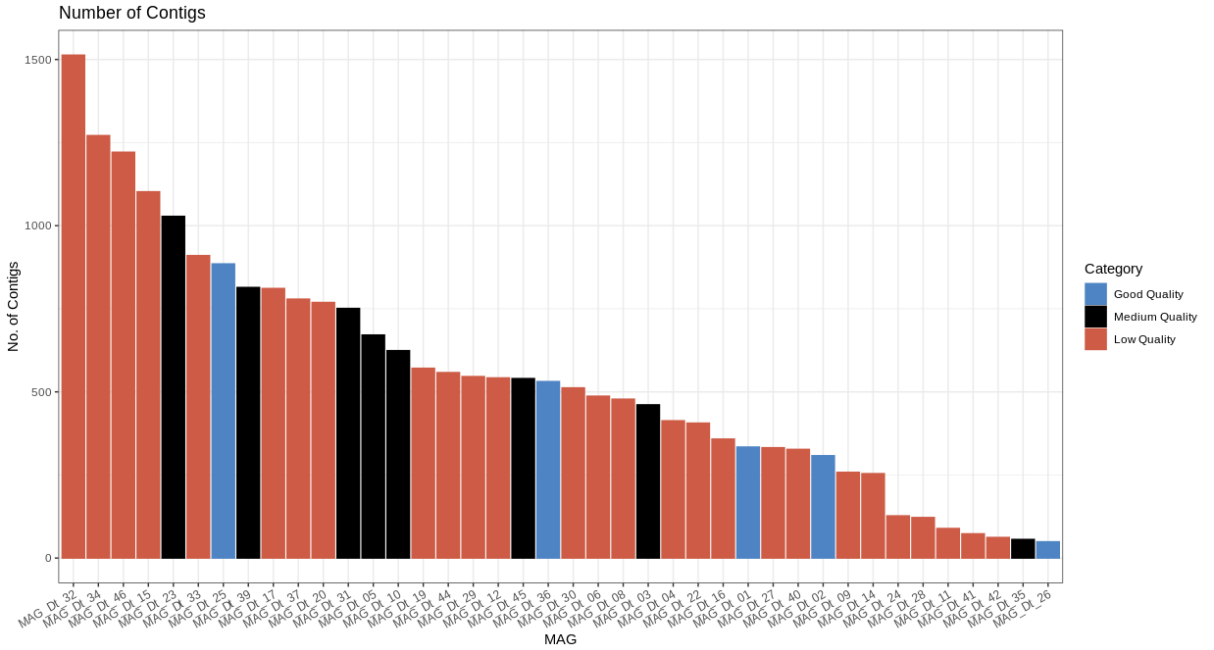


Figure 15. Number of contigs each MAG consists of, classified by quality. Tendencies can be observed that high and medium quality MAGs consist of fewer contigs as their low-quality counterparts.

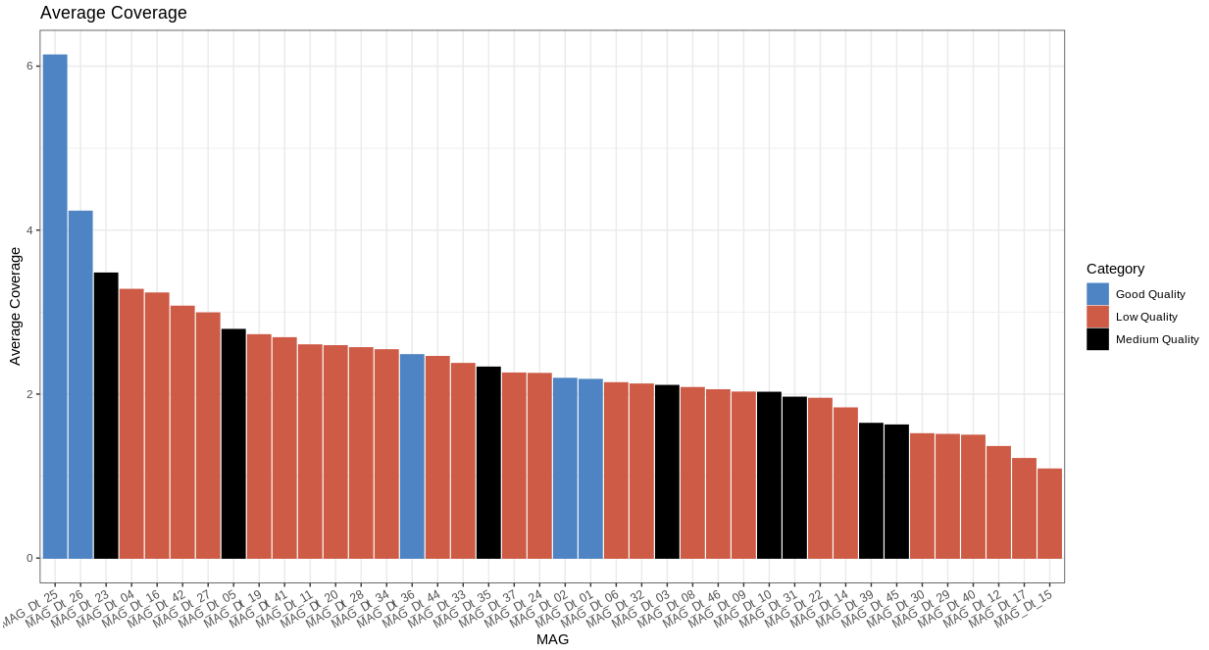


Figure 16. Average coverage of all contigs for each MAG, categorized by quality. Although the average coverage is relatively low in all MAGs, high quality MAGs seem to consist of higher average coverages than medium and low-quality MAGs.

4.2 Phylogeny and Taxonomy

The results from all four approaches were considered for taxonomic classification, as far as it was possible to obtain reasonable results.

4.2.1 Ribosomal Protein Analysis

The abundance of the different ribosomal proteins is displayed in the Figure 17. Due to the low amount of 19 equal representatives in 37 MAGs, it was not possible to draw further conclusions from this approach.

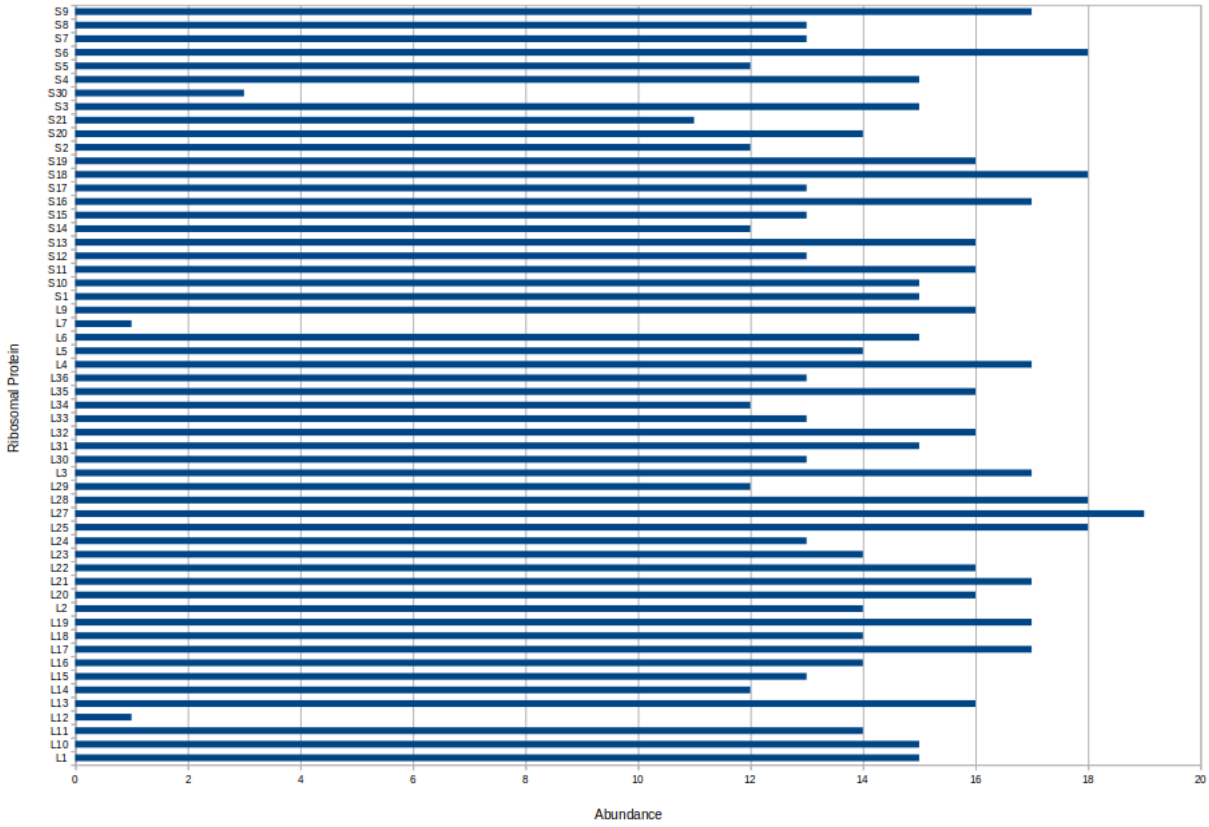


Figure 17. Abundance of the extracted ribosomal proteins, sorted by type. Due to too few representatives, it was not possible to continue with a phylogenetic analysis based on this approach

4.2.2 Lowest Common Ancestor

A taxonomic table for all MAGs, except unclustered MAGs, was obtained by comparison with the non-redundant database (NCBI). At least all MAGs could be classified into certain phyla. Seven high or medium quality MAGs could be characterized more specifically. An overview of the taxonomic results is shown in Table 3., where the corresponding phylum and the most specific classification were displayed for each MAG, classified by Autometa. [67]

4.2.3 Average Nucleotide Identity

For all MAGs, the highest ANI-score was evaluated and plotted in Figure 18. Only one MAG, MAG_Dt_03, obtained an ANI-score over 80%. The remaining MAGs had average nucleotide identities between 74% to 79% (A). Due to the low similarity scores, no further classification of the MAGs could be made. Furthermore, a heatmap based on all ANI-scores was generated for each MAG compared to the reference genomes. (B)

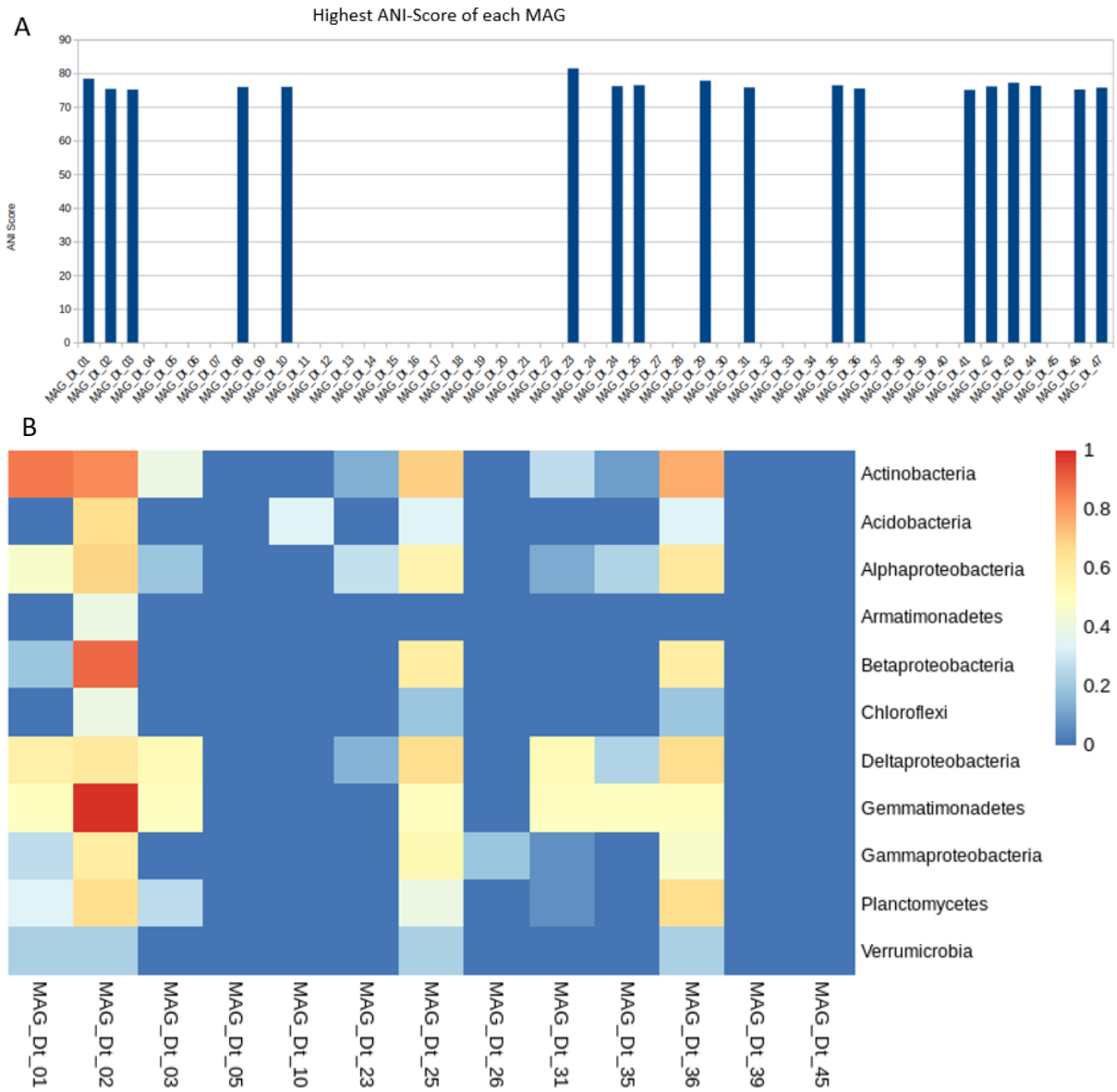


Figure 18. Results from FastANI [97] analysis: (A) the highest ANI-scores for each MAG, regardless the quality. If no bar can be seen, the ANI score is below 75%. (B) Relative abundance of obtained ANI-scores for high- and medium quality MAGs, with regards to the corresponding phylum. I.e. ANI scores over 75% could be observed between all Gemmatimonadetes and MAG_Dt_02. Therefore, the relative abundance is equal to 1.

From the relative abundance of ANI scores, MAGs could be classified into two subgroups: ones, that show similarities with a wide range of organisms in different phyla (group 1), and others, which have only similarity to reference organism of one or no phylum (group 2). The results were summarized in Table 2.

Table 2.: Summary of most likely taxonomic classifications based on highest ANI scores and the relative abundance of ANI scores per phylum.

MAG	CLASSIFICATION	MOST ABUNDANT REFERENCE PHYLUM	QUALTIY CLASSIFICATION
MAG_Dt_01	Group 1	*Actinobacteria*	High
MAG_Dt_02	Group 1	*Gemmatimonadetes*	High
MAG_Dt_03	Group 1	*Deltaproteobacteria*	Medium
MAG_Dt_05	Group 2	-	Medium
MAG_Dt_10	Group 2	*Acidobacteria*	Medium
MAG_Dt_23	Group 1	*Alphaproteobacteria*	Medium
MAG_Dt_25	Group 1	*Actinobacteria*	High
MAG_Dt_26	Group 2	*Gammaproteobacteria*	High
MAG_Dt_31	Group 1	*Deltaproteobacteria*	Medium
MAG_Dt_35	Group 1	*Gemmatimonadetes*	Medium
MAG_Dt_36	Group 1	*Actinobacteria*	High
MAG_Dt_39	Group 2	-	Medium
MAG_Dt_45	Group 2	-	Medium

4.2.4 Phylogenetic Tree

An unrooted (Figure 19.) and a rooted (Figure 20.) phylogenetic tree were generated based on the results from OrthoFinder. [98] In both trees, the phyla containing a MAG were coloured and described, whereas all other phyla were not marked. From the rooted tree, the phylogenetic distance and the node values can be determined, for further evaluation of the reliability.

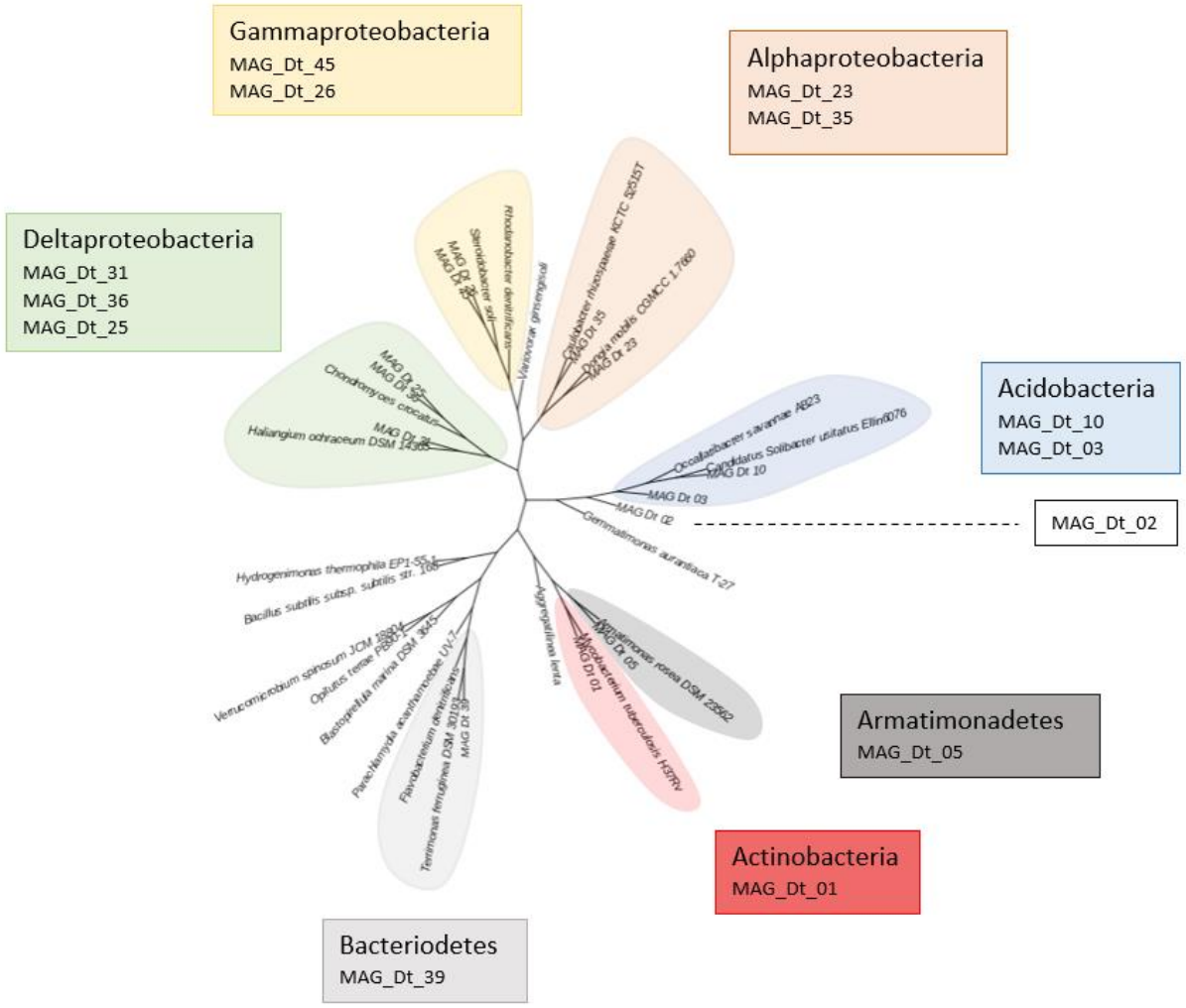


Figure 19. Unrooted tree, based on whole proteomes from the displayed organisms. Fewer representatives were used from each phylum to ensure readability. All MAGs could be assigned at least to a phylum, except MAG_Dt_02, which was identified between *Acidobacteria* and *Gemmatimonadetes*.

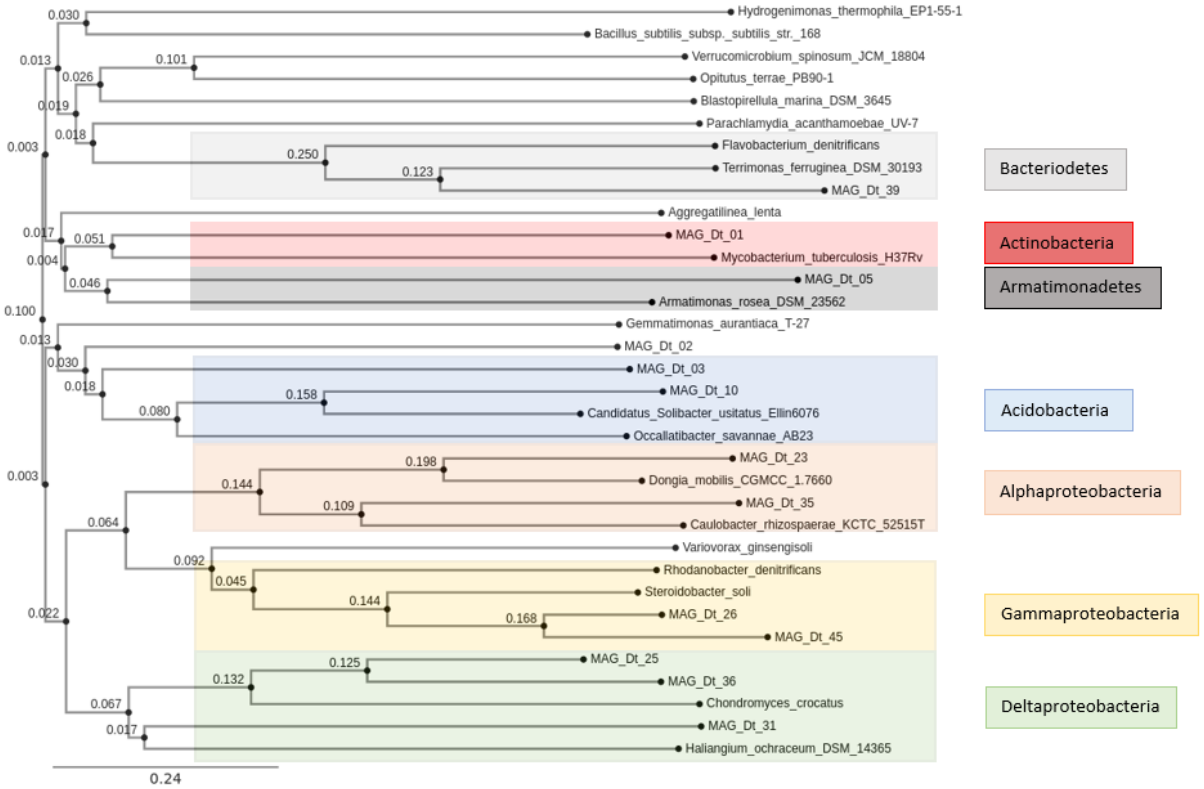


Figure 20. Rooted tree based on the whole proteomes from the displayed organisms. In contrast to the unrooted tree, phylogenetic relationships and distances can be observed. The node values are calculated by the software tool OrthoFinder [98] by combining phylogenetic trees from all orthogroups. Therefore, higher values indicate that organisms share a higher number of orthologues.

Although MAG_Dt_02 was determined to be part of the *Acidobacteria*, it could not be defined exactly. When using another set of reference genomes, MAG_Dt_02 was characterized as a *Gemmatimonadetum*.

4.2.5 Taxonomic Classification

The results from the three approaches for taxonomic classification and the final description are summarized in Table 3. The classification into a phylum is based on results from OrthoFinder, [98] Autometa, [67] and ANI. [97] The most accurate classification results from the Autometa taxonomic table. [67] MAG_Dt_39 and MAG_Dt_45 were only classified by the results from OrthoFinder, [98] due to the lack of taxonomic classification by Autometa. [67]

Table 3.: Final taxonomic classification of high- and medium-quality MAGs, based on the results from OrthoFinder [98] and Autometa. [67]

MAG	PHYLUM	MOST ACCURATE CLASSIFICATION	CLASSIFICATION
MAG_Dt_01	*Actinobacteria*	-	High
MAG_Dt_02	*Acidobacteria**	-	High
MAG_Dt_03	*Acidobacteria*	-	Medium
MAG_Dt_05	*Armatimonadetes*	-	Medium
MAG_Dt_10	*Acidobacteria*	*Solibacteriaceae (Family)*	Medium
MAG_Dt_23	*Alphaproteobacteria*	*Rhodospirillaceae (Family)*	Medium
MAG_Dt_25	*Deltaproteobacteria*	*Myxococcales (Order)*	High
MAG_Dt_26	*Gammaproteobacteria*	*Steroidobacter (Genus)*	High
MAG_Dt_31	*Deltaproteobacteria*	*Rhodospirillaceae (Family)*	Medium
MAG_Dt_35	*Alphaproteobacteria*	-	Medium
MAG_Dt_36	*Deltaproteobacteria*	*Myxococcales (Order)*	High
MAG_Dt_39	*Bacteriodetes*	-	Medium
MAG_Dt_45	*Gammaproteobacteria*	*Steroidobacter (Genus)*	Medium

Due to the classification of MAG_Dt_02 as *Acidobacterium* using the Autometa [67] classification, it was possible to define the MAG as an *Acidobacterium* despite the inaccurate results from OrtohFinder.

The classified phyla were used to calculate the Shannon-Index to indicate the bacterial diversity of the MAGs in the metagenome. The higher the calculated Shannon-Index is, the higher is the diversity in a certain sample. The Shannon-Index of the high and medium quality MAGs resulted in a value of 1.84, compared to the maximum value of 2.65 for 13 species. Including also low-quality MAGs, the value increased to 2.87, in comparison to a maximum value of 3.61. Regarding low-quality MAGs, two more phyla, involving the *PVC-group* and *Gemmatimonadetes*, were covered.

4.3 Postprocessing Analysis

4.3.1 Gene Prediction and Annotation

The predicted proteins were annotated using the KEGG [107] and PANNZER2 (UniProt) [108] database. The absolute amount of predicted proteins and annotated proteins from KEGG [107] and PANNZER2 [108] were counted and displayed in Figure 21. MAGs with high quality consisted of more proteins than MAGs with medium quality, which could be expected due to the lower completeness level of the latter ones. In all MAGs more proteins were annotated through the PANNZER2 [108] database than through the KEGG database. [107] The highest annotation percentage was observed in MAG_Dt_01 (75%, PANNZER2 [108]), whereas in two MAGs (MAG_Dt_25 and MAG_Dt_05) fewer than 50% of the proteins were characterized. For protein annotation of metabolic pathways and BGCs, annotations from the KEGG database [107] were taken into account first. All unannotated proteins were then characterized using the PANNZER2 [108] annotation. Overall proteins, more than 50% could not be annotated completely.

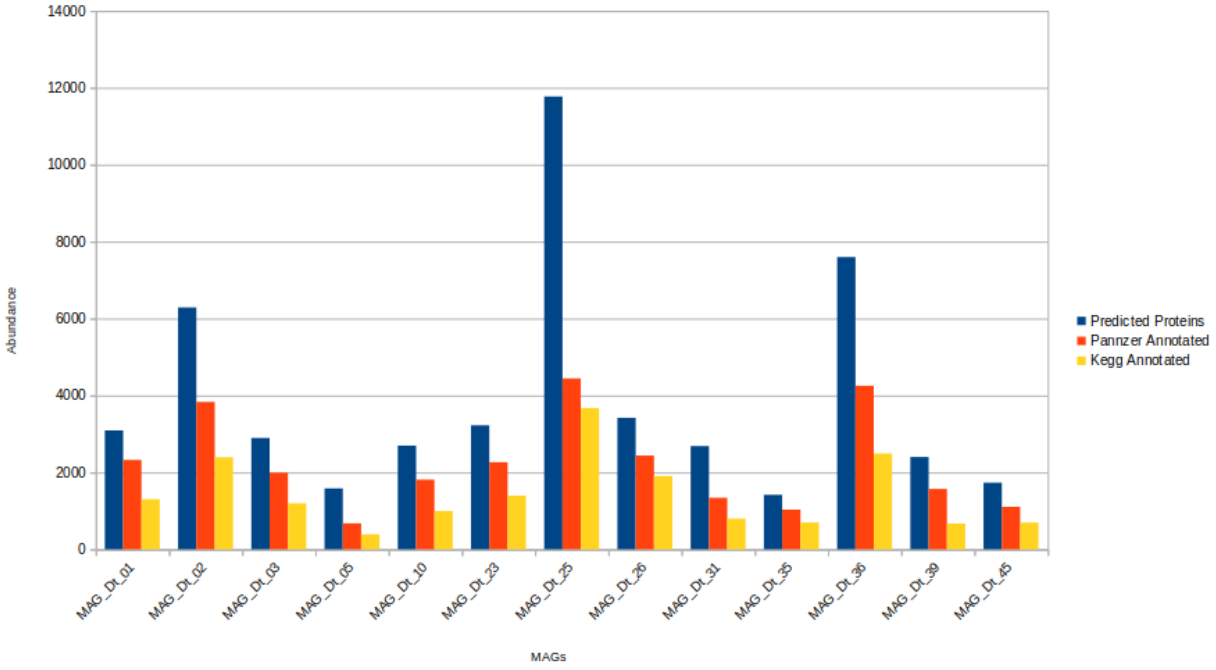


Figure 21. Absolute abundance of predicted (blue), and annotated (PANNZER2 [108] – orange, KEGG [107] – yellow) proteins in high- and medium-quality MAGs.

4.3.2 Metabolic Pathway Analysis

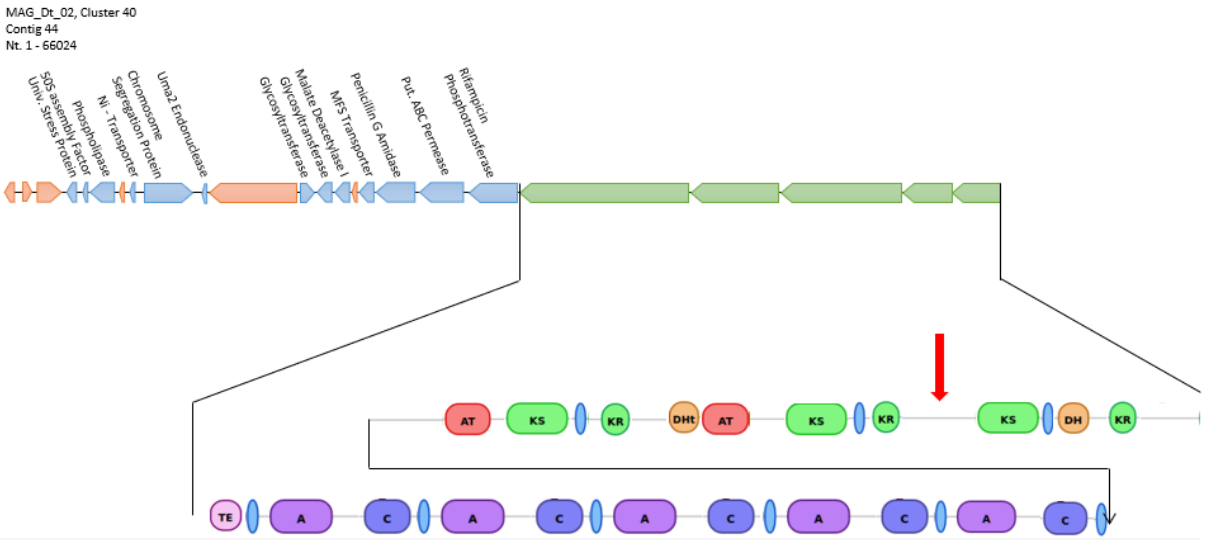
4.3.2.1 Enediyne Biosynthesis

The pathway for 9-membered enediyne core molecules is completely present in MAG_Dt_02. Although previous studies showed these core enzymes cluster together, it could be discovered, that only the enzymes *E5* (contig 105, gene 37, nt. 771-1667), *pksE* (contig 105, gene 38, nt. 1664-7321), and the thioesterase *E10* (contig 105, gene 39, nt. 7318-7749) are grouped within one contig, whereas enzymes *E3* (contig 27, gene 2, nt. 710-1732) and *E4* are located on other sites in the genomes. Interestingly, 13 different peptides are characterized as enediyne biosynthesis enzyme *E4*, whereas all other copies are only present once. Five *E4* enzymes are located on one contig (contig 22, genes 19, 20, 22, 23, 26). Besides the core enzymes another uncharacterized enzyme involved in enediyne biosynthesis, *E7* (contig 45, gene 31, nt. 47388-48761), is contained. The enzymes *E5*, *pksE*, and *E10* are followed by three uncharacterized enzymes and a DNA ligase. Despite the annotated *pksE*, the contig was not detected by antiSMASH, [85] but was discovered by KEGG [107] and PANNZER2 [108] annotation instead.

4.3.2.2 Quorum Sensing

Three MAGs were identified to potentially synthesize molecules involved in QS. In MAG_Dt_35 and MAG_Dt_39 parts of the biosynthetic pathway for autoinducer-2 (AI-2) were observed. The core enzyme for the final release of homocysteine, *LuxS* (contig 6304, gene 3, nt. 1437-1844), was identified in MAG_Dt_39 as well as the methyltransferase (contig 10691, gene 2, nt. 1663-2346). The only missing enzyme, S-adenosyl homocysteine nucleosidase (contig 6, gene 53, nt. 53980-54657), to replace the adenosine by a ribosyl group, was only identified in MAG_Dt_35, whereas the core enzyme, *LuxS*, could not be found in MAG_Dt_35.

In MAG_Dt_02 the core enzyme for the biosynthesis of the autoinducer-1 (AI-1) was annotated as Penicillin Amidase G. After further investigation of the two enzymes, it was revealed that both proteins were also characterized as Acyl-homoserine-lactone synthases (contig 44, gene 17, nt. 24036-26432). Additionally, one of the core enzymes was identified within a hybrid NRPS-PKS BGC, displayed in Figure 22. Furthermore, an MFS transporter and an ABC permease were found next to the Acyl-homoserine-lactone synthase, which are potentially involved in the export. The modules from the NRPS-PKS enzyme were characterized as incomplete due to a missing starting module and a missing acyltransferase domain, indicated with a red arrow in Figure 22. The other enzyme, essential for the synthesis of the precursor S-adenosyl-L-methionine, was also identified as S-adenosylmethionine synthase (contig 1, gene 121, nt. 142347-143528)



Legend: MAG_Dt_02, Cluster 40, Contig 44, Nt. 1 – 66024;
Domain: A = Adenylation, C = Condensation, KS = Ketosynthase, AT = Acyltransferase,
KR = Ketoreductase, DH = Dehydratase, TE = Thioesterase, DHt = Dehydratase (variant)

Figure 22. A BGC containing an almost complete hybride NRPS-PKS enzyme complex, as well as the biosynthetic core enzyme for autoinducer-1 biosynthesis, annotated as Penicillin G amidase. The red arrow indicates the position of the missing Acyltransferase (AT) domain. Additionally, the starting module is missing due to the cut off of the contig. The genes are divided into core (green), annotated (blue) and unannotated (red) genes.

4.3.2.3 Alkanesulfone Metabolism

In MAG_Dt_01 the enzyme phosphosulfolactate phosphatase (contig 217, gene 3, nt. 2604-3293) was discovered which is an essential part of the coenzyme M biosynthesis. Another gene potentially involved in the biosynthetic pathway was identified on a surrounding contig as the NAD+ - dependent L-2-Hydroxycarboxylate Dehydrogenase (contig 215, genes 2 and 3, nt. 331-1722 and nt. 1719-2408). Further genes located nearby were characterized to contain 4 Fe – 4 S catalytic regions (peroxiredoxin, contig 218, gene 6, nt. 3783-4214), which were described to catalyze the decarboxylation and reduction steps [84] of the alkanesulfone precursor or with potential sulfate transfer activity. As a sulfide source two potential enzymes were identified: the cysteine desulfurylase (contig 220, gene 1, nt. 2-1330), which is part of the thiamine metabolism, and the polysulfide reductase (contig 53, gene 1, nt. 2-718). Besides the putative biosynthetic enzymes, an alkanesulfone importer subunit (*SsuB*) was detected, which would explain the missing phosphosulfolactate synthase, though, the other subunits of the importer complex (*SsuA* and *SsuC*) were not found. This enzyme was identified within MAG_Dt_39 (contig 5221, gene 2, nt. 826-1599). Although all other enzymes of the coenzyme M biosynthetic pathway could not be found in MAG_Dt_39, the biosynthetic pathway could be covered by a symbiotic biosynthesis, as shown in Figure 23, of MAG_Dt_01 and MAG_Dt_39, where only the final decarboxylating enzyme was not identified.

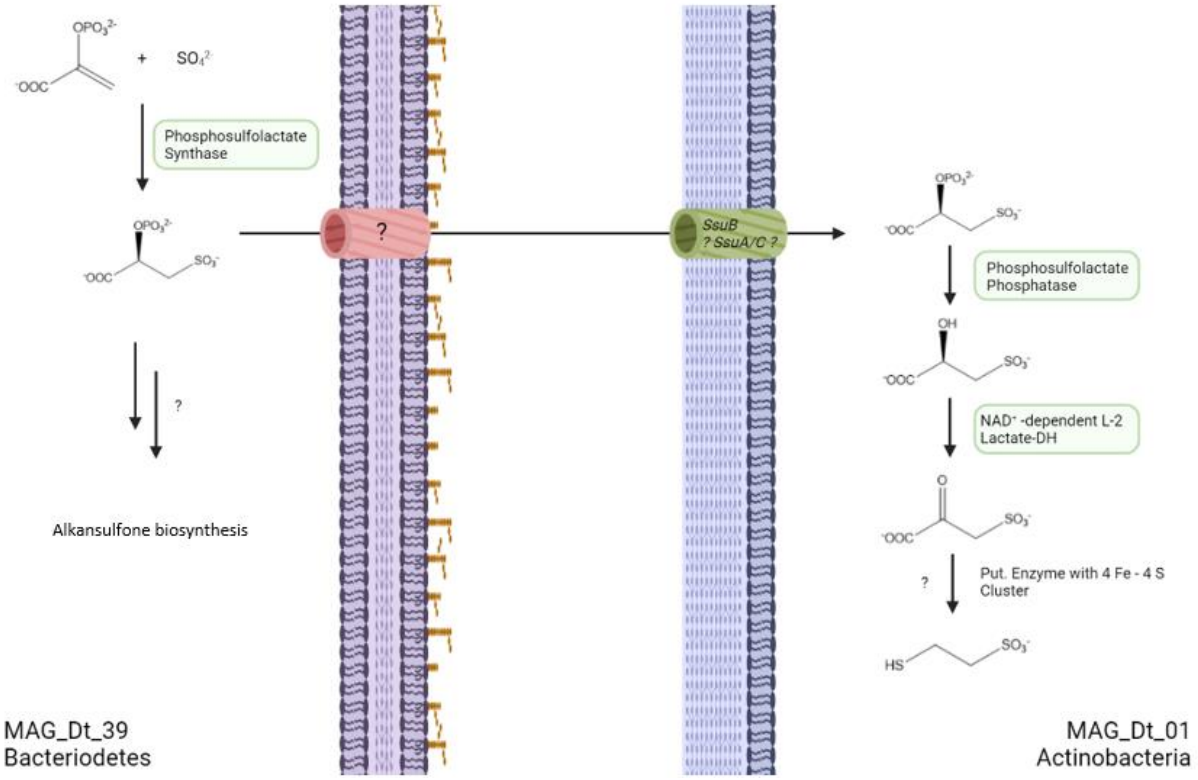


Figure 23. Graphic representation of the identified enzymes for putative alkansulfone biosynthesis in MAG_Dt_39 and MAG_Dt_01. (The graphic was created on the Biorender-platform)

In both MAGs a contamination value below 5% was recorded, whereas the completeness levels of MAG_Dt_39 (33%) and MAG_Dt_01 (85%) differed significantly. Therefore, it must be considered that the whole biosynthetic pathway could occur in MAG_Dt_39.

4.3.3 Biosynthetic Gene Cluster Analysis and Secondary Metabolite Potential

A total of 1230 BGCs were predicted within the rhizobiome, and 424 BGCs within the characterized MAGs, regardless the quality. The distribution of the BGC types is shown in Figure 24. The most abundant BGCs in both, MAGs (blue) and the rhizobiome (orange), are classified as “Fatty Acid”, “Saccharide”, “Putative” or “Other” gene clusters. BGCs from the first two categories are assumed to be involved in the primary metabolism and are therefore not further analyzed. The undefined BGCs (“Putative” and “Other”) are a diverse group, which is not further classified so far. Therefore, they could synthesize bioactive substances, potentially, but are difficult to characterize. The most important gene cluster types for potential bioactive SM are grouped as “Terpenes”, “NRPS” or “PKS” – containing as well as “Bacteriocins” and “Lantipeptides”.

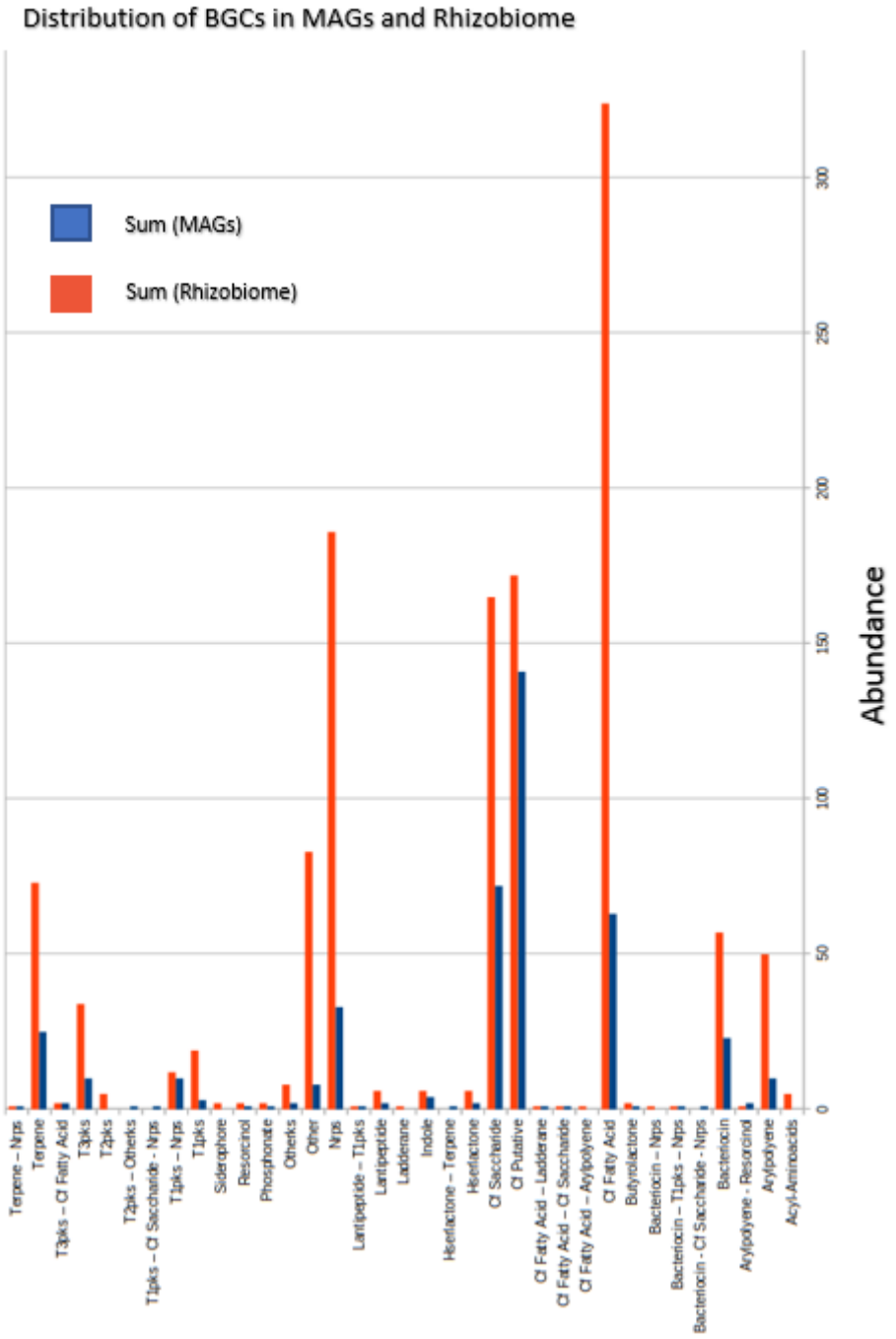


Figure 24. Graphic representation of predicted BGCs, sorted by type and if they were observed in MAGs or the rhizobiome.

The BGCs, containing “NRPS” or “PKS” modules are further grouped by the completeness of the predicted modules. (Figure 25.) As expected, the amount of BGCs, overall and complete, is higher in the rhizobiome compared to MAGs due to the larger dataset. An exception is the T2PKS-OtherKS and one T1PKS-NRPS hybrid gene cluster, which was found in MAGs and was not predicted in the rhizobiome. This can be explained by the missing reassembly step with SPAdes [65] of the rhizobiome-contigs. Therefore, the gene cluster is split into two gene clusters in the rhizobiome. Furthermore, the complete amount of “NRPS” – BGCs (186) is not shown in Figure 25 to better display the differences between the remaining clusters. Most of the complete gene clusters can be associated with certain MAGs. An exception are the six complete “NRPS” gene clusters, which must be part of not assembled microorganisms from the rhizobiome. Overall, 13 complete BGCs, containing “PKS”, “NRPS” or hybrids were obtained and annotated.

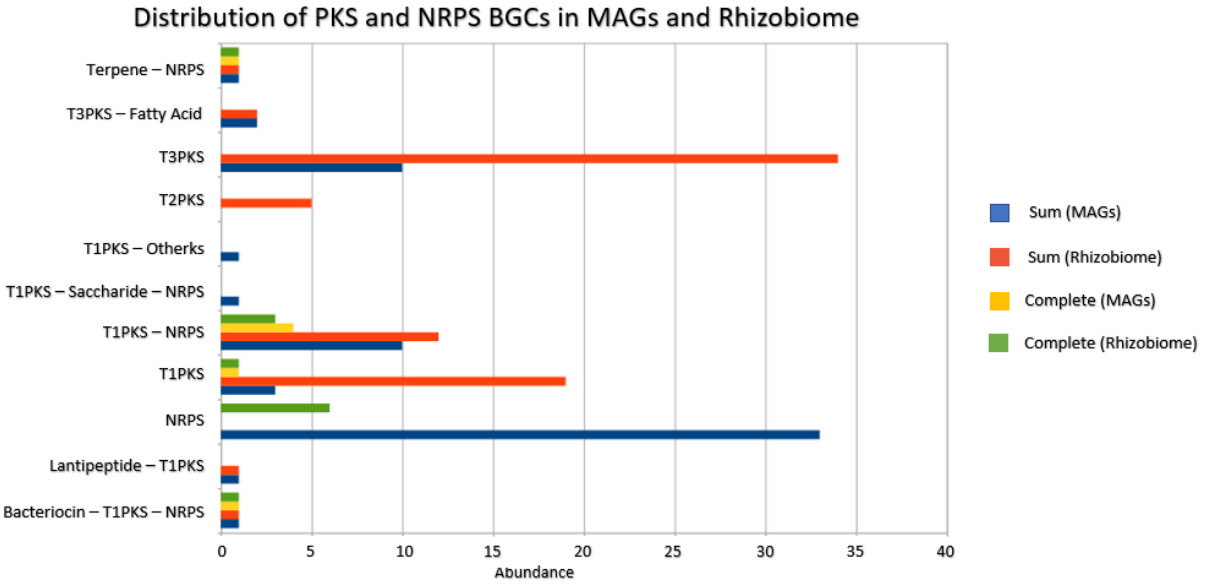


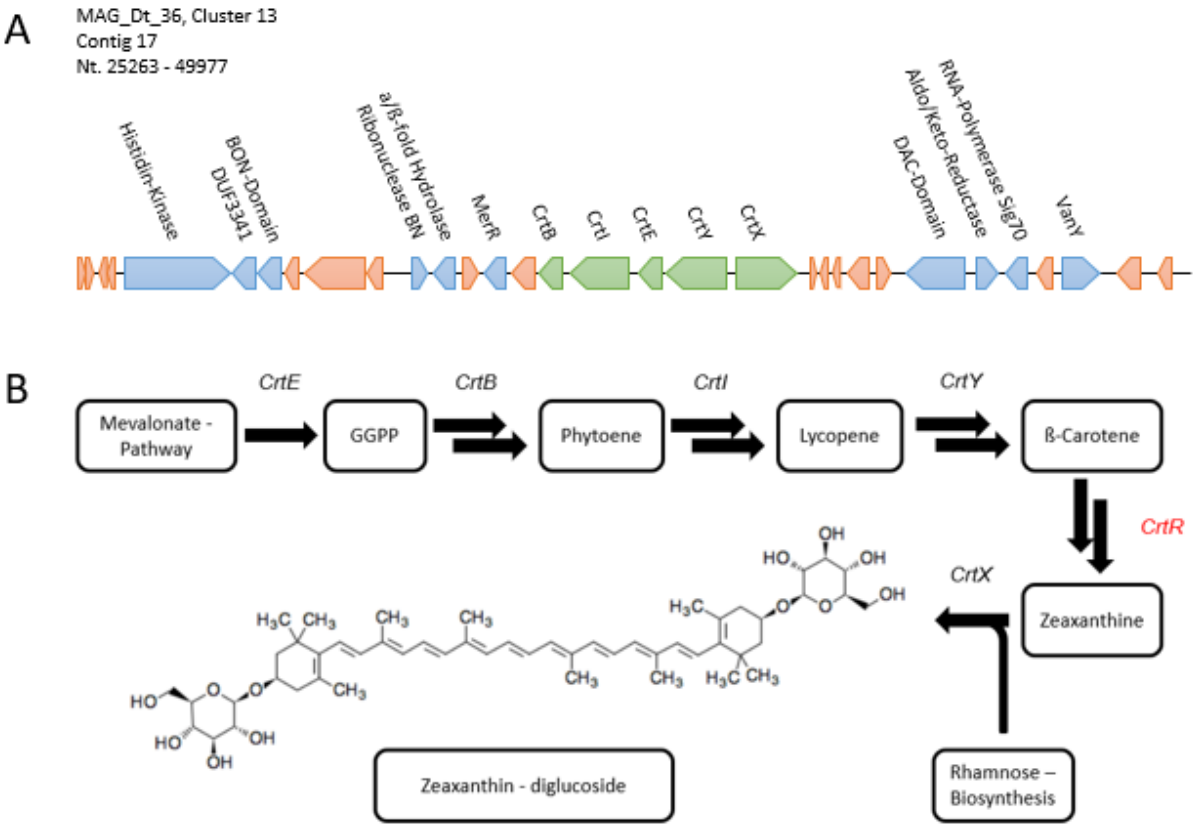
Figure 25. Graphic representation of predicted BGCs containing PKS and/or NRPS. The BGCs were sorted by completeness (yellow and green for complete modules) and if they were observed in MAGs or in the rhizobiome.

In Figure 36. the characterized BGCs are shown for each MAG. High-quality MAGs contain the most BGCs, ranging from 19 (MAG_Dt_26) to 71 (MAG_Dt_02), followed by less distributed MAGs with medium quality, covering only one (MAG_Dt_05) up to 24 BGCs (MAG_Dt_03). Some low-quality MAGs contain no BGCs, whereas MAG_Dt_44 contains as much as some high (MAG_Dt_01) or medium quality MAGs (MAG_Dt_03).

All BGCs, which produce potential bioactive SM are described in the following chapters. The core genes were marked in green, further annotated genes were coloured in blue and undefined genes were displayed in red. Further classifications for certain genes were described in the corresponding chapter.

4.3.3.1 Terpene

The analyzed terpene cluster was found in MAG_Dt_36 on contig 17 with the core genes 35-39 (gene 35: nt. 35263-36213, gene 36: nt. 36210-37727, gene 37: nt. 37720-38799, gene 38: nt. 38796-41260, gene 39: 40040-41260). In Figure 26. the BGC for Zeaxanthin-diglucoside biosynthesis is displayed, including the core enzymes (A), as well as the pathway (B), according to the KEGG-database. [107] The isopentenyl-subunits, synthesized through the mevalonate pathway, could be concatenated to GGPP (Geranyl-geranyl-pyrophosphate) by *CrtE* (GGPP-synthase). After dehydration by *CrtB* (15-cis phytoene synthase), terminal cyclization by *CrtI* (phytoene desaturase), and by *CrtY* (lycopene ß-cyclase), and reduction catalyzed by *CrtR* (Cytochrome P450 reductase), which was not identified in this BGC, but within the MAG. Finally, D-rhamnose subunits could be added by *CrtX* (Zeaxanthin-glycosyltransferase).



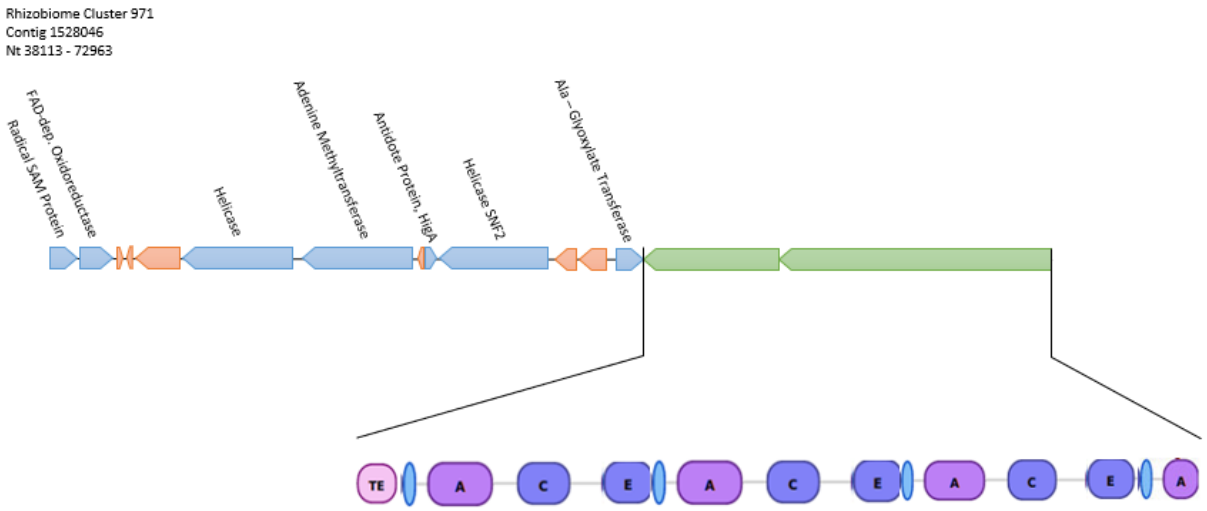
Legend: MAG_Dt_36, Cluster 13, Contig 17, Nt. 25263 – 49977

Figure 26. (A) A BGC, containing all enzymes, necessary for Zeaxanthin – diglucoside biosynthesis, except *CrtR*. (B) Reconstructed biosynthetic pathway starting from the Mevalonate pathway, according to KEGG-database. [107] *CrtE* synthesises GGPP from Isopentenyl-subunits, followed by dehydration (*CrtB*), desaturation (*CrtI*), terminal cyclization (*CrtY*) and glycosyl-transferase (*CrtR*). The genes are divided into core (green), annotated (blue) and unannotated (red) genes.

4.3.3.2 *Non-ribosomal Protein Synthetase*

The annotated NRPS-BGC was found in the rhizobiome, on contig 1528046 with the core genes 41 (nt. 58113-61736) and 42 (nt. 61791-72962) and is displayed in Figure 27. In the described gene cluster, two core enzymes were identified at the end of the contig, which, in turn, could be interpreted as a complete gene cluster, but also as an incomplete BGC. The putative missing sequence could be lost during the assembly step. In the displayed modules the obligate condensation (C) and adenylation (A) regions were defined together with an epimerization domain (E).

Besides the NRPS the protein *HigA* (gene 36, nt. 50938-51222), an antidote protein, was identified in the BGC. The condensation modules were predicted to use L – amino acids as substrates followed by an epimerization. In the second annotated gene cluster (Figure 38.) the core enzymes for non-ribosomal protein synthesis were not found at the terminal sites of the BGC. Hence, the completeness of the core enzyme was assumed, although a second condensation domain in the sixth module was observed.

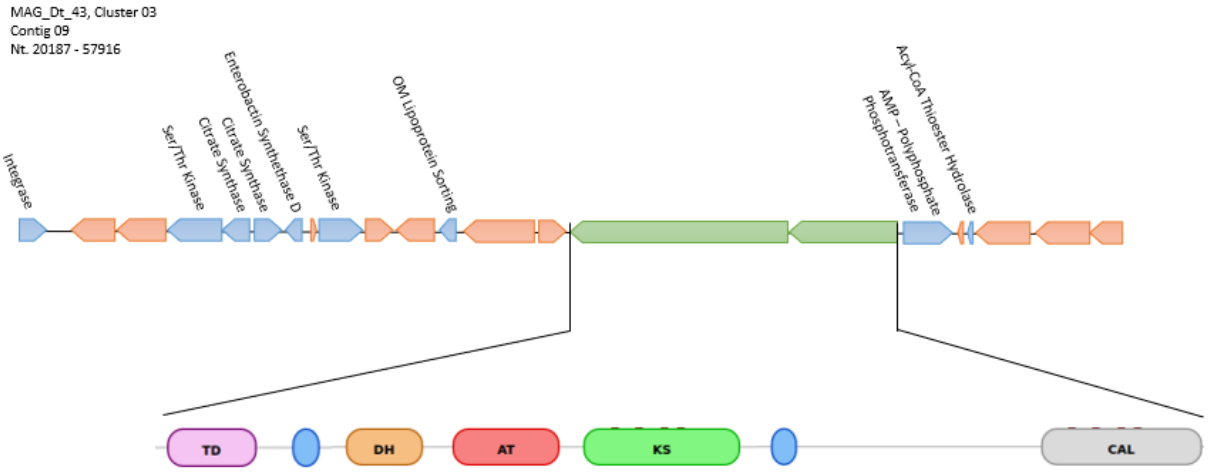


Legend: Rhizobiome, Cluster 971, Contig 1528046, Nt. 38113 – 72963;
Domain: A = Adenylation, C = Condensation, E = Epimerization, TE = Thioesterase

Figure 27. NRPS-containing BGC. Although the modules seem to be complete, it would be possible that the enzyme was disrupted by the termination of the contig. The genes are divided into core (green), annotated (blue) and unannotated (red) genes.

4.3.3.3 *Type 1 Polyketide Synthase*

The only complete type 1 PKS was identified as single modular, containing a ketosynthase (KS), an acyltransferase domain (AT), and a dehydratase domain (DH), in MAG_Dt_43, on contig 9. Furthermore, the enterobactin synthetase D subunit (*EntD,* contig 9, gene 24, nt. 30064-30753), was annotated upstream to the core enzyme. The diphosphatase *EntD* was already reported to act as a supporting enzyme on the core enzyme *EntB*, the substrate carrier enzyme, in the biosynthetic pathway of enterobactin. Nevertheless, enterobactin and derivates were known to be produced by NRPS instead of type 1 PKS. [101] Other subunits for enterobactin biosynthesis (*EntA-C* and *EntE-F*) could not be identified in the MAG, due to the contamination value of over 100%.



Legend: MAG_Dt_43, Cluster 03, Contig 09, Nt. 20187 – 57916; Core genes 32 (nt. 40187-45382) and 33 (nt. 45379-49113)
Domain: KS = Ketosynthase, AT = Acyltransferase, DH = Dehydratase, TD = Terminal Reductase, CAL = Coenzyme A Ligase

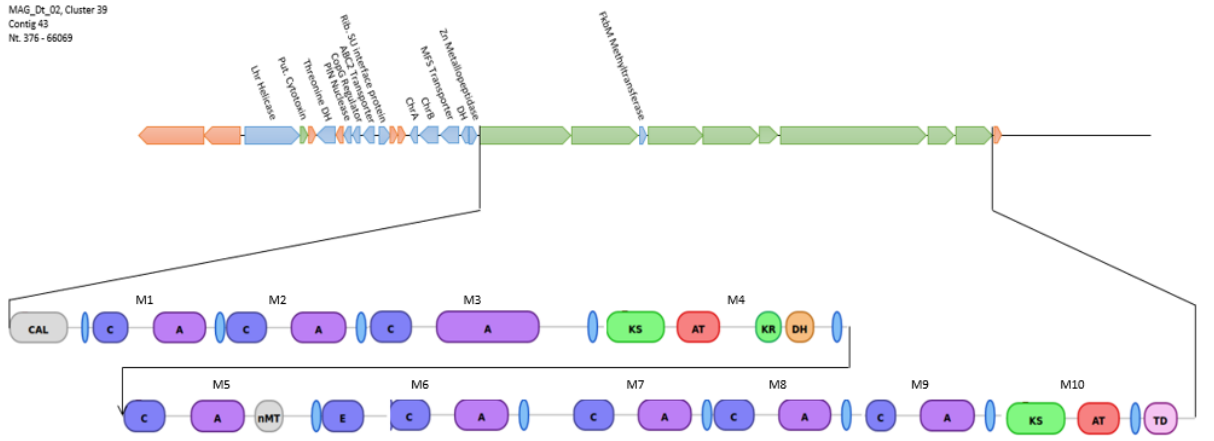
Figure 28. A complete type-1 PKS-BGC, observed in the excluded MAG_Dt_43. The genes are divided into core (green), annotated (blue) and unannotated (red) genes.

4.3.3.4 NRPS-PKS-Hybrids

All observed and annotated hybrid NRPS-PKS gene clusters were identified with consecutive modules due to missing termination modules within the core enzymes (see Supplementary). Furthermore, all parts of the core enzyme showed the same orientation on the BGC, which was taken as further evidence that the modules could be part of one core enzyme. Only in cluster 28 from MAG_Dt_02, the core genes were observed to be counter orientated. (Figure 39.)

The displayed BGC was found in MAG_Dt_02 on contig 43 with the core genes 19-27 (gene 19: nt. 20376-28442, gene 20: nt. 28462-33798, gene 21: nt. 33795-34724, gene 22: nt. 23721-40123, gene 23: nt. 40120-46092, gene 24: nt. 46096-48015, gene 25: nt. 48012-58535, gene 26: nt. 58510-61575, gene 27: nt. 61572-65792). Besides the starting (CAL) and termination module (TD), ten modules were characterized, from which eight were predicted to be NRPS-modules and two PKS-modules. The integrated methyltransferase (nMT) was observed in module five, whereas another methyltransferase was predicted as a single enzyme (*FkbM*). Methyltransferases from the *FkbM* family were described to contain two well-conserved regions as well as a variable amino acid chain in between. Motamedi, H., et. al., identified a member of the *FkbM* family as part of the biosynthesis of immunosuppressant's. [102] In contrast to six NRP-modules containing possible condensation domains for L-amino acids and an adenylation domain (M1, M2, M3, M7, M8, M9), in one module (M6) an epimerization domain, followed by a D-amino acid condensation domain was characterized. In addition to the obligate ketosynthase (KS) and acyltransferase domain (AT), a ketoreductase- (KR) and a dehydratase domain (DH) were identified in module M4, in contrast to module M10.

Furthermore, a putative cytotoxic peptide (contig 43, gene 4, nt. 10616-10843) was identified but could not be specified in more detail, as well as two unspecified transporter proteins, Type-2 ABC and MFS were identified upstream to the core enzyme.



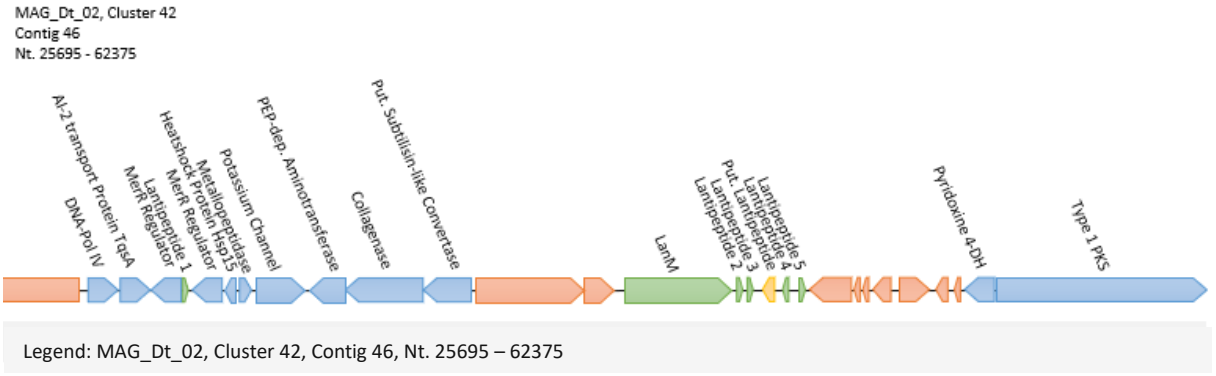
Legend: MAG_Dt_02, Cluster 39, Contig 43, Nt. 376 – 66069;
Domain: A = Adenylation, C = Condensation, KS = Ketosynthase, AT = Acyltransferase, E = Epimerization, KR = Ketoreductase, DH = Dehydratase, TD = Terminal Reductase, CAL = Coenzyme A Ligase, nMT = Nitrogen-Methyltransferase
Module: NRPS = M1, M2, M3, M5, M6, M7, M8, M9; PKS = M4, M10

Figure 29. Hybrid NRPS-PKS BGC containing complete modules and a separately annotated methylation domain within the core enzyme. The cluster contains complete modules (M1-10) and is not located at a terminus of the contig, which means that no modules were cut off. The BGC contains 8 NRPS modules and 2 PKS modules. The genes are divided into core (green), annotated (blue) and unannotated (red) genes.

4.3.3.5 *Lantipeptide*

The BGC with five lantipeptides was predicted in the high-quality MAG_Dt_02, on contig 46 with the core gene 37 (nt. 45695-48982) and Lantipeptide 1 (nt. 31007-31134), Lantipeptide 2 (nt. 48979-49191), Lantipeptide 3 (nt. 49210-49322), Lantipeptide 4 (nt. 49745-49842), Lantipeptide 5 (nt. 50019-50149) and Put. Lantipeptide (nt. 49352-49711). The core enzyme within this cluster was defined as *LanM* enzyme. All putative lantipeptides were observed next to the *LanM* enzyme except Lantipeptide 1, which was identified at the beginning of the BGC. Nearby the Lantipeptide 1 a gene was annotated as "AI-2 transporter protein *TqsA*" for the transport of Autoinducer-2, a molecule involved in QS. Hetzberg, M., and colleagues described *TqsA* as a transporter and regulator, due to expression changes, of autoinducer-2 molecules and, hence, of biofilm formation. [103]

In contrast to the lack of enzymes involved in AI-2 biosynthesis in MAG_Dt_02, the AI-2 sensor kinase *LuxQ* was detected in the neighbouring contig (MAG_Dt_02, contig 45). In addition to the *LuxQ* sensor kinase (contig 45, gene 2, nt. 452-2215), the two-component repressor and regulation factors, *LuxO* (contig 7, gene 51, nt. 62205-63590) and *LuxR* (contig 22, gene 38, nt. 52808-53467), were identified within the genome. In former studies, [104] AI-2 was described to inhibit transcription repressor proteins to enable transcription of certain genes. Furthermore, it was shown by Wang, L., and colleagues, that an accumulation of AI-2 in the cytoplasm increased the transcription of sensor and transporter enzymes, such as *TqsA* transporters. [104] Due to the close location, it would be possible that the Lantipeptide 1 would be coexpressed with the *TqsA* (contig 46, gene 26, nt. 28916-30040) gene.



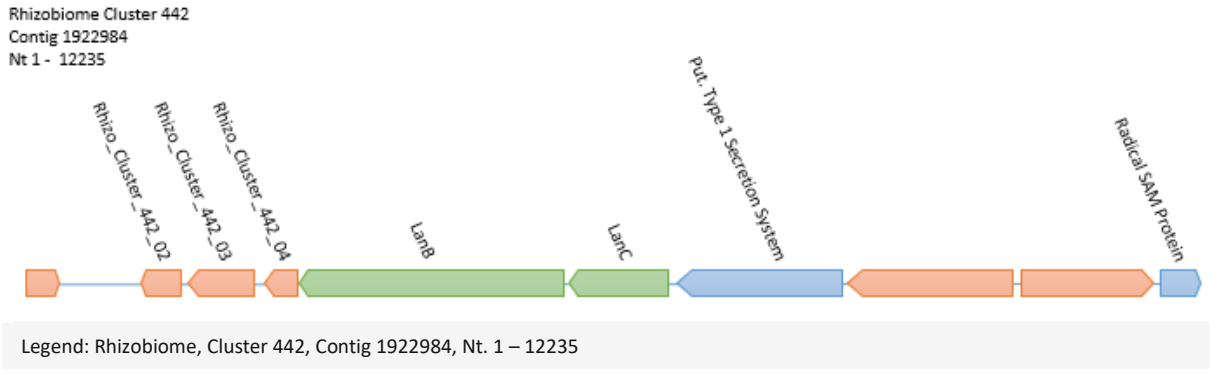
Legend: MAG_Dt_02, Cluster 42, Contig 46, Nt. 25695 – 62375

Figure 30. Type II lantipeptide containing BGC. Four of five predicted BGCs are located near the core enzyme *LanM*, whereas the last is relatively far away, but clusters together with an AI-2 transporter protein, which is necessary to import QS signals. The genes are divided into core (green), annotated (blue) and unannotated (red) genes. One unannotated peptide, located within four lantipeptides was marked (yellow) and investigated.

Table 4.: Overview of the five predicted lantipeptides, separated by leader and core peptide as well as the unannotated small peptide, between the lantipeptides. Furthermore, all potential serine (yellow), threonine (blue) and cysteine (red) residues were marked.

LANTIPEPTIDE	LEADER PEPTIDE	CORE PEPTIDE
Lantipeptide 1	MKLTRYGQHIDKRRSSGYSGDSTSLWG	GDhbDhbEDhbLRDhaRDhaVFLCL
Lantipeptide 2	MKKKIDVARAWRDEEYYLGLTEEERASLGA	HPDhaGLIEVDGDhaLLKDhbVVGGVA DhbLVDDhbCDhaAICDhbPCPPRQCY
Lantipeptide 3	LREAFRVSRKPSPKGRAGRAPAEARPLLAG	RFCIDCQ
Lantipeptide 4	LSKLLSKNVLTWPETRGIILA	VVRECGDhbPPVP
Lantipeptide 5	VARSLAAA	AQGRAWAARGRAADhbARGRA VADGDRAAAARCRAGHARLLAAAANP
Put. Lantipeptide	MPEITSFADYVTDWERLLAAVANNEAGLPDLGPQRTSLEDILEEAKAVSTRQDASRSQLS ADAKRRREILFEGRAAASRLRAALKGHFGGHNEKLVEFGARPIRQRRTAKLVDPPLAVEX	

After further investigation of the leader sequences, which were expected to be conserved, [50] only the leader sequence of Lantipeptide 2 showed a sequence similarity of 66,67% with other leader sequences in the non-redundant protein database (NCBI). The other leader sequences could not be characterized. Finally, the sequence of the uncharacterized protein between Lantipeptide 3 and Lantipeptide 4, marked in yellow, was extracted and all serine, threonine, and cysteine residues were marked in the corresponding colours.



Legend: Rhizobiome, Cluster 442, Contig 1922984, Nt. 1 – 12235

Figure 31. A type-I lantipeptide cluster, containing the core enzymes *LanB* and *LanC*. Despite no lantipeptides could be predicted, the three short peptides downstream to the core enzymes were investigated further. The genes are divided into core (green), annotated (blue) and unannotated (red) genes.

Another potential BGC synthesizing lantipeptides was found in the rhizobiome on contig 1922984 with the core genes 5-6 (gene 5: nt. 2157-5444, gene 6: nt. 5441-6745). In contrast to the prior described cluster containing *LanM* as core enzyme, the two enzymes *LanB* and *LanC* were characterized. The three following genes could not be annotated and no leader and core – sequences for putative lantipeptides were predicted. The multiple sequence alignment with 20 defined lantipeptides (from [50]), as well as the lantipeptides from the prior described BGC, gave no significant results. Therefore, the sequences were screened for potential serine (yellow), threonine (blue), and cysteine (red) residues, which were marked in Table 5.

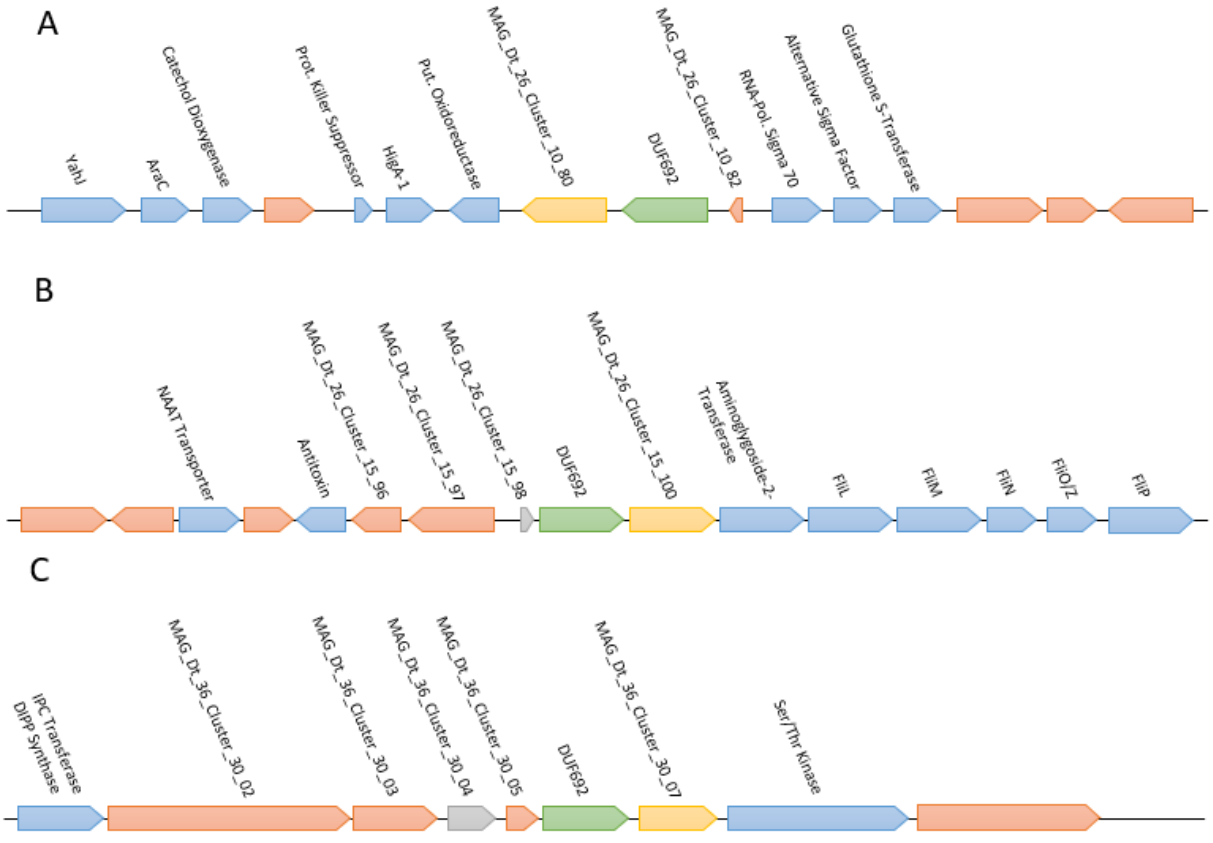
Table 5.: Amino acid sequence of the three short peptides, downstream to the core enzymes. Serine, threonine and cysteine residues were coloured as before to obtain an impression for potential post-translational modifications.

LANTIPEPTIDE	SEQUENCES
Rhizo_Cluster_442_02	MDDKARIAEIEHRIMAAFAAGDAEALVAQYTEDAVLLSPDYPAIQGRAAILEA YRAALDEYEMRLETVVEETEVAGDWAWMRGRFEHTSTRKADGAATTARGK YLVIARRDPDGAWRFHRDAFNLDEPRTX
Rhizo_Cluster_442_03	MRFYTQQECDEWLSDRQRTKPDAAPGVHRERISYPPEPYRIFSVAHWMATSL TYRMPALLWVTEWGIWPSSESWHLYYKLRQAYQDQRLLHEAPGHLFLEHEAE DLASFLQVAMLNGWGGYLLTQADYVNAFFSHDEYIDFFAEREEALADVRTEL GKSGTAEX
Rhizo_Cluster_442_04	MKKLQKKLSLNRETLRNLSGHELQGIVGGVTGTCCNSSTETGDSCVTCNVTTHC TTTNYCTQGACYTDLCX

It was observed that the first two lantipeptides (Rhizo_Cluster_442_02 and Rhizo_Cluster_442_03) did not contain any cysteine residues, but several serine and threonine residues instead. In contrast, up to eight cysteine residues were detected in the putative core region at the end of the peptide sequence. Due to the high variability of the leader sequences, [50] it was not possible to divide the amino acid sequences into leader and core regions. However, in all sequences were serine, threonine, or cysteine residues observed in the last third of the sequences.

4.3.3.6 Bacteriocin

Bacteriocin clusters can be divided into two types: ones are characterized by proteins, containing the DUF692 domain (Domain of unknown function) and those, which encode for specific and already known bacteriocins, e.g. maritimacin. In all 16 BGCs, coding for a DUF692-domain-containing protein, several other genes could also not be annotated. Besides the unknown genes surrounding the DUF692 gene, oxidoreductases and silver efflux pumps are located near the core gene in four BGCs and could be involved in the biological function, consequently. In Figure 32. three of the 16 BGCs are represented.



Legend: A) MAG_Dt_26, Cluster 10, Contig 06, Nt. 91713 – 102567
B) MAG_Dt_26, Cluster 15, Contig 08, Nt. 98709 – 109566
C) MAG_Dt_36, Cluster 30, Contig 106, Nt. 1 – 9968

Figure 32. Three representative BGCs containing the DUF692-core protein. The unannotated (red) and the DUF692-domain containing (green) proteins were used for the multiple sequence to investigate potential similarities. Further annotated genes were marked in blue. Genes in yellow or grey were discovered to cluster together after the multiple sequence alignment. Grey genes could potentially be silver efflux pumps due to the fact that they cluster together with an annotated silver efflux pump.

Due to the lack of further information about the neighbouring proteins in literature, [55] a multiple sequence alignment of all undefined genes was performed and displayed in Figure 33 B, including the corresponding sequence identity in %. The genes were renamed from X1 to X69 as described in Table 7.

From the multiple sequence alignment and the corresponding cladogram (Figure 33 A), the undefined genes could be divided into three main groups: the highest sequence identity was shared within the DUF692 containing proteins, which could be expected due to the only member with at least one already defined domain. Two other groups of genes were defined by the clusters in Figure 33. In the first group of genes (X24-X31), with higher sequence identities, genes were located before the core gene and were marked in grey in the corresponding gene clusters. The gene X28 (MAG_Dt_10, cluster 6, gene 7) was already annotated as a silver efflux pump. The other three BGCs, also containing silver efflux pumps, showed the same pattern. The second group of genes (X1-X16) shared a lower sequence identity and were observed to be located after the DUF692-domain-containing protein in 14 of 16 cases. Further attempts by comparing these genes to defined surrounding genes, i.e. oxidoreductase or hydrolase, were unsuccessful. Due to the relatively low sequence identity, no domains in all genes were observed and could therefore not be further investigated. The annotated BGCs were displayed as described in Figure 43 to Figure 46.

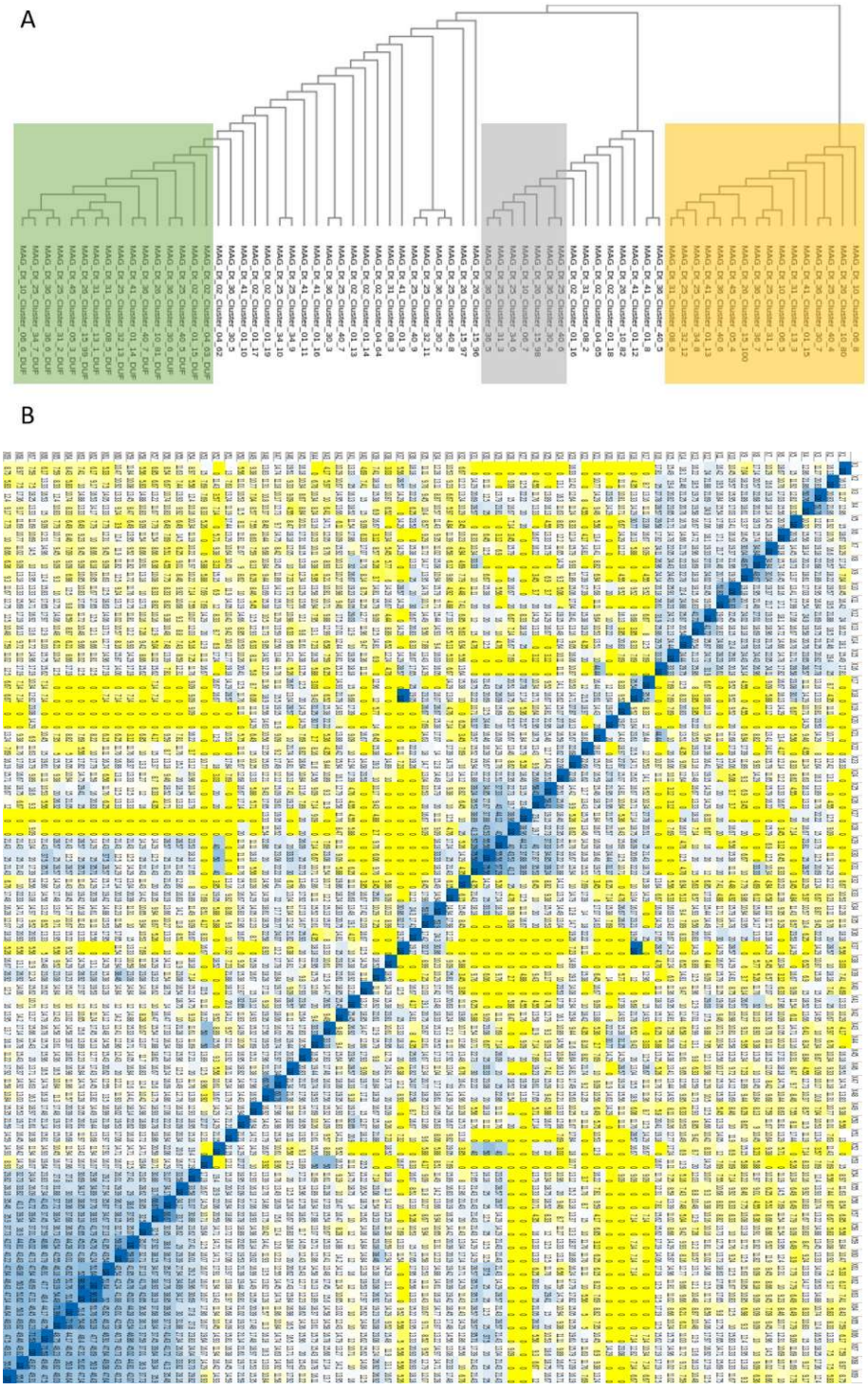


Figure 33. Results from the multiple sequence alignment using unnotated peptides from DUF692-containing BGCs. (A) Similarity based tree, including coloring for the three clusters, observed in (B). The clusters from the heatmap (B) are above the marked clusters in the dendrogram (A). Overall three clusters could be identified: DUF692-Domain containing proteins (green), putative silver efflux pumps (grey) and a cluster containing only unannotated proteins (yellow).

5 Discussion

The shotgun metagenomics sequencing of the rhizobiome from *D. traunsteineri* resulted in five high and eight medium quality MAGs, which were characterized by their taxonomy. In addition, the metabolic potential of the MAGs and the rhizobiome as well as 1230 BGCs from which 420 could be assigned to MAGs, to produce potential bioactive secondary metabolites were analyzed and annotated.

The introduction of next-generation sequencing techniques and their application to microbiomes from different sources resulted in a tremendous increase of microbial diversity. Several studies examined that the phylogenetic diversity in soil ranges from 10^2 to 10^6 different taxa dependent from hosts, environmental factors (i.e. drought, temperature, pH, etc.) or depth. [109] In many habitats the most abundant discovered phyla were *Acidobacteria, Proteobacteria* or *Bacteriodetes*. [110] [111] Kaur and Sharma summarized metagenomic studies investigating root associated bacteria from orchids worldwide and discovered that they are dominated by *Proteobacteria*, followed by *Actinobacteria*. [121] In the present study the two most abundant phyla in high or medium quality MAGs were *Proteobacteria* (7) and *Acidobacteria* (3). In our study we could discover a MAG, MAG_Dt_02, which was identified as *Acidobacterium* or *Gemmatimonadetes*, dependent on the approach and the reference organisms used. Therefore, it is important to define the taxonomy based on several approaches to obtain robust results. Additionally, it must be considered that shotgun metagenomic experiments result in significantly fewer discovered MAGs in contrast to diversity studies using targeted metagenomic analyses. [113] Together with the general minority of eukaroytic or archaea specific sequences compared to bacteria, [114] it was only possible to generate bacterial MAGs. This, in turn, prevents comparison with former studies, mainly focussing on fungal communities in the rhizosphere of orchids. [122] Furthermore, the investigation of BGCs of the rhizobiome associated with orchids has been completely overlooked so far, though, it has been shown that SM impact hosts and its microbiome. [14] Due to the high potential of SM and their influence on environments and hosts, the authors highly recommend consideration in future studies.

Besides the microbial diversity, numerous shotgun metagenomic experiments discovered the diversity of BGCs in different habitats. Similar to the microbial diversity, the BGC types also vary significantly. [115] [116] [117] The most often identified groups are NRPS, PKS, terpenes, bacteriocins or others, which are not further classified. Although, the abundance is an important value for description of a microbial community, studies often lack a detailed investigation and annotation of BGCs, in particular of underrepresented BGC types such as lantipeptides. Despite the higher effort of BGC mining, it could be worth since lantipeptides cover a novel class of potential antimicrobials with putative activity against biofilms, which are often not destroyed by other antibiotic compounds. [51] In the present study, we could identify five lantipeptides with predicted leader and core sequence, from which one could be regulated by a quorum sensing dependent mechanism, and four putative lantipeptides, which are located downstream to the corresponding core enzymes on the BGC. Furthermore, BGCs characterized as bacteriocins must be distinguished since a huge amount of them is predicted based on the DUF692-domain containing protein by antiSMASH. Due to the still unknown function of this protein [55] it should be classified in another way to avoid wrong conclusions regarding the antimicrobial potential. In addition to the current studies to identify the biological function of the DUF692-domain containing protein, it should be considered that surrounding genes are not annotated as well and could play a role in their bioactivity. Therefore, it was tried to identify sequence similarities of the unknown neighbouring genes by a multiple sequence alignment. Although it was not possible to identify similar core regions in most of the genes, we were able to estimate the function of six undefined genes as potential silver efflux pumps, which supports the findings that the core protein is often located near silver efflux pumps. Another

cluster of genes, sharing a low sequence similarity was identified. These genes were found as the neighbouring gene (downstream) of the core protein in 13 out of 16 BGCs. The investigation of the DUF692-domain containing protein and the neighbouring genes is important since it has been shown that they are potentially involved in the biosynthesis of 3-thiaglutamate, which is known to inhibit jasmonate and ethylene signalling pathways in plants. [55]

Besides the investigation of predicted BGCs by software tools such as antiSMASH, it is worth to further examine genomes with focus on PKS or NRPS. Especially enzymes, which are poorly characterized so far, such as the enediyne PKS core enzyme *pksE* could probably not be predicted. [118] Furthermore, it should be considered that genes for biosynthesis of secondary metabolites are not restricted to BGCs. To overcome this problem, genome wide co-expression analyses could be used to investigate enzymes which are necessary for biosynthesis. [57]

In addition to the detailed investigation and annotation of BGCs within high quality MAGs, the BGC prediction for the metagenome should not be omitted since potential BGCs from organisms which were not reconstructed would be lost. As an example, the only complete type 1 PKS was identified in MAG_Dt_43 (Figure 28), with a contamination value above 100%.

Besides BGC analysis, the very specific enzyme phosphosulfolactate phosphatase, which is involved in coenzyme M biosynthesis, [84] was identified in MAG_Dt_01, a high-quality MAG, characterized as an *Actinobacterium*. Although no further enzymes for coenzyme M biosynthesis could be identified, potential proteins, which could substitute their function were found in the MAG, including a lactate-dehydrogenase and a sulfurylase for sulfate activation. The only enzyme containing a 4Fe-4S cluster, which is thought to catalyze the final sulfidation and decarboxylation step, [84] within the investigated contigs, was annotated as peroxiredoxin. This enzyme plays a major role in oxidative stress response by inhibition of H2O2. Additionally, it has been investigated that pyruvate (and derivates potentially also) can be decarboxylated in presence of H2O2. The combined knowledge would describe a putative new pathway for coenzyme M biosynthesis as a side product of oxidative stress response. Although the hypothesis needs to be proven, a concerted reaction of decarboxylation in presence of H2O2 followed by a sulfidation by peroxiredoxin could be an explanation for the difficulty [84] of characterizing a single enzyme to perform the decarboxylation and sulfidation step.

6 Conclusion

With the present study we could show that high and medium quality MAGs can be obtained from the rhizosphere of *D. traunsteineri* based on a shotgun metagenomics approach despite the high complexity of the rhizobiome. Furthermore, we could use the created MAGs, as well as the unbinned metagenomic contigs to predict 1230 BGCs using antiSMASH. In combination with functional gene annotation based on KEGG and PANNZER2 databases, we identified and described BGCs potentially producing bioactive SM. These are the first BGCs described from the rhizobiome of *D. traunsteineri*.

7 References

[1] Paun O., Bateman R., Fay M., Hedren M., Civeyrel L., Chase M., Stable Epigenetic Effects Impact Adaption in Allopolyploid Orchids (Dactolyrhiza: Orchidaceae), Molecular Biology and Evolution, Volume 27, Issue 11, 2010, doi.org/10.1093/molbev/msq150

[2] Pillon Y, Fay MF, Hedrén M, Bateman RM, Devey DS, Shipunov AB, van der Bank M, Chase MW, Evolution and temporal diversification of western European polyploid species complexes in Dactylorhiza (Orchidaceae), Taxon, 2007, vol. 56, DOI:10.2307/25065911

[3] Paun O, Fay MF, Soltis DE, Chase MW, Genetic and epigenetic alterations after hybridization and genome doubling. Taxon. 2007 Aug; 56(3):649-56, PMCID: PMC2980832

[4] http://burgenlandflora.at/pflanzenart/dactylorhiza-traunsteineri/, 02.05.2021

[5] http://www.aho-bayern.de/taxa/da_trau.html, 05.05.2021

[6] https://www.infoflora.ch/de/flora/dactylorhiza-traunsteineri.html, 05.05.2021

[7] Siles, J.A., Margesin, R. Seasonal soil microbial responses are limited to changes in functionality at two Alpine forest sites differing in altitude and vegetation. Sci Rep 7, 2204 (2017). https://doi.org/10.1038/s41598-017-02363-2

[8] Stott P., How climate change affects extreme weather events, Science, 2016, DOI: 10.1126/science.aaf7271

[9] Najera F., Dippold M., Boy J., Seguel O., Koester M., Stock S., Merino C., Kuzyakov Y., Matus F., Effects of drying/rewetting on soil aggregate dynamics and implications for organic matter turnover, Biology and Fertility of Soils, 2020, vol. 56, DOI:10.1007/s00374-020-01469-6

[10] Montgomery and Biklé, The Hidden Half of Nature: The Microbial, 2016

[11] Hartmann T., From waste products to ecochemicals: fifty years research of plant secondary metabolism, Phytochemistry, 2007, Vol. 68, Iss. 22-24, DOI: 10.1016/j.phytochem.2007.09.017

[12] Ruiz B., Chavez A., Forero A., Garcia-Huante Y., Romero A., Sanchez M., Rocha D., Sanchez B., Rodriguez-Sanoja R., Sanchez S., Langley E., Production of microbial secondary metabolites: regulation by the carbon source, Critical Reviews in Microbiology, 2010, doi: 10.3109/10408410903489576

[13] O'Brien J., Wright G., An ecological perspective of microbial secondary metabolites, Current Opinion in Biotechnology, 2011, doi: 10.1016/j.copbio.2011.03.010

[14] Jacoby R., Koprivova A., Kopriva S., Pinpointing secondary metabolites that shape the composition and function of the plant microbiome, Journal of Experimental Botany, 2021, 72(1):57-69, doi: 10.1093/jxb/eraa424

[15] Woese CR Microbiol Rev. 1987 Jun; 51(2):221-71, PMCID: PMC373105

[16] Hugenholtz P., Goebel B., Pace N., Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity, Journal of Bacteriology, 1998, 180(18): 4765-4774, PMCID: PMC107498

[17] Wade W., Unculturable bacteria – the uncharacterized organisms that cause oral infections, Journal of the Royal Society of Medicine, 2002, 95(2):81-83, doi: 10.1258/jrsm.95.2.81

[18] Coker J., Recent advances in understanding extremophiles, F1000 Faculty, 2019, doi: 10.12688/f1000research.20765.1

[19] Hamady, M. & Knight, R. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. Genome Res. 19, 1141–1152 (2009).

[20] Haas B., et. al., Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons, Genome Research, 2001, doi: 10.1101/gr.112730.110

[21] Shieh F., et. al., ChimericSeq: An open-source, user-friendly interface for analyzing NGS data to identify and characterize viral-host chimeric sequences, PLOS ONE, 2017, doi.org/10.1371/journal.pone.0182843

[22] Callahan B., et. al., DADA2: High-resolution sample interface from illumine amplicon data, Nature Methods, 2016, 13(7):581-3, doi: 10.1038/nmeth.3869
[23] Quince, C., Walker, A., Simpson, J. et al. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35, 833–844 (2017). https://doi.org/10.1038/nbt.3935
[24] Medema, M., Kottmann, R., Yilmaz, P. et al. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol 11, 625–631 (2015). https://doi.org/10.1038/nchembio.1890
[25] Schwecke T, Aparicio JF, Molnár I, et al. The biosynthetic gene cluster for the polyketide immunosuppressant rapamycin. Proc Natl Acad Sci U S A. 1995;92(17):7839-7843. doi:10.1073/pnas.92.17.7839
[26] Liu, X. F., Xiang, L., Zhou, Q., Carralot, J.-P., Prunotto, M., Niederfellner, G., et al. (2016). Actinomycin D enhances killing of cancer cells by immunotoxin RG7787 through activation of the extrinsic pathway of apoptosis. Proc. Natl. Acad. Sci. U.S.A. 113, 10666–10671. doi: 10.1073/pnas.1611481113
[27] Newman, D. J., and Cragg, G. M. (2016). Natural products as sources of new drugs from 1981 to 2014. J. Nat. Prod. 79, 629–661. doi: 10.1021/acs.jnatprod.5b01055
[28] https://mibig.secondarymetabolites.org/stats, 19.08.2021
[29] Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, van Santen JA, Tracanna V, Suarez Duran HG, Pascal Andreu V, Selem-Mojica N, Alanjary M, Robinson SL, Lund G, Epstein SC, Sisto AC, Charkoudian LK, Collemare J, Linington RG, Weber T, Medema MH. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res. 2020 Jan 8;48(D1):D454-D458. doi: 10.1093/nar/gkz882
[30] Epstein, S.C., Charkoudian, L.K. & Medema, M.H. A standardized workflow for submitting data to the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository: prospects for research-based educational experiences. Stand in Genomic Sci 13, 16 (2018). https://doi.org/10.1186/s40793-018-0318-y
[31] Le Govic Y., et. al., Non-ribosomal Peptide Snythetase Gene Clusters in Human Pathogenic Fungus Scedosporium apiospermum, Fronties in Microbiology, 2019, https://doi.org/10.3389/fmicb.2019.02062
[32] Izoré, T., Candace Ho, Y.T., Kaczmarski, J.A. et al. Structures of a non-ribosomal peptide synthetase condensation domain suggest the basis of substrate selectivity. Nat Commun 12, 2511 (2021). https://doi.org/10.1038/s41467-021-22623-0
[33] Miller B., Gulick A., Structural Biology of non-ribosomal peptide synthetases, Methods in Molecular Biology, 2016, 1401: 3-29, doi: 10.1007/978-1-4939-3375-4_1
[34] Agrawal S., et. al., Nonribosomal peptides from marine microbes and their antimicrobial and anticancer potential, Frontiers in Pharmacology, 2017, https://doi.org/10.3389/fphar.2017.00828
[35] Challis G., Naismith J., Structural aspects of non-ribosomal peptide biosynthesis, Curr Opin Struct Biol, 2004, 14(6): 748-756, doi: 10.1016/j.sbi.2004.10.005
[36] Caboche S, Leclere V, Pupin M, Kucherov G, Jacques P (2010) Diversity of monomers in non-ribosomal peptides: towards the prediction of origin and biological activity. J Bacteriol 192: 5143–5150, doi: 10.1128/JB.00315-10
[37] Hwang S., Lee N., Cho S., Palsson B., Cho B., Repurposing Modular Polyketide Synthases and non-ribosomal peptide synthetases for novel chemical biosynthesis, Front. Mol. Biosci., 2020, https://doi.org/10.3389/fmolb.2020.00087
[38] Dutta S., Whicher J., Hansen D., Hale W., Chemler J., Congdon G., Narayan A., Häkansson K., Sherman D., Smith J., Skiniotis G., Structure of a modular polyketide synthase, Nature, 2014, 510(7506):512-7, doi: 10.1038/nature13423

[39] Shen B., Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms, Current Opinion in Chemical Biology, 2003, 7(2):285-95, doi: 10.1016/s1367-5931(03)00020-6
[40] Liu F., Garneau S., Walsh C., Hybrid Nonribosomal peptide-Polyketide interfaces in epothilone biosynthesis: minimal requirements at N and C termini of EpoB for elongation, CellPress, 2004, Vol. 11, doi: 10.1016/j.chembiol.2004.08.017
[41] Mizuno C., Kimes N., Lopez-Perez M., Auso E., Rodriguez-Valera F., Ghai R., A Hybrid NRPS-PKS gene cluster related to the bleomycin family of antitumor antibiotics in Alteromonas macleodii strains, PLOS ONE, 2013, doi: 10.1371/journal.pone.0076021
[42] Gatto G., McLoughlin S., Kelleher N., Walsh C., Elucidating the substrate specificity and condensation domain activity of FkbP, the FK520 pipecolate-incorporating enzyme, Biochemistry, 2005, 44(16):5993-6002, DOI:10.1021/bi050230w
[43] Tang G., Cheng Y., Shen B., Leinamycin biosynthesis revealing unprecedented architectural complexity for a hybrid polyketide synthase and nonribosomal peptide synthetase, Chem. Biol., 2004, 11(1):33-45, doi: 10.1016/j.chembiol.2003.12.014
[44] Nielsen, M., Isbrandt T., Petersen L., Mortensen U., Andersen M., Hoof J., Larsen T., Linker Fleixibility Facilitates Module Exchange in Fungal PKS-NRPS Engineering, PLOS ONE, 2016, doi.: 10.1371/journal.pone.0161199
[45] Macek, B., Forchhammer, K., Hardouin, J. et al. Protein post-translational modifications in bacteria. Nat Rev Microbiol 17, 651–664 (2019). https://doi.org/10.1038/s41579-019-0243-0
[46] Mayer, H.; Bauer, H.; Breuss, J.; Ziegler, S.; Prohaska, R. Characterization of rat LANCL1, a novel member of the lanthionine synthetase C-like protein family, highly expressed in testis and brain. Gene 2001, 269 (1−2), 73−80, doi: 10.1016/s0378-1119(01)00463-2
[47] Park, S.; James, C. D. Lanthionine synthetase components C-like 2 increases cellular sensitivity to adriamycin by decreasing the expression of P-glycoprotein through a transcription-mediated mechanism. Cancer Res. 2003, 63 (3), 723−727, PMID: 12566319
[48] Zhang, Q.; Doroghazi, J. R.; Zhao, X.; Walker, M. C.; van der Donk, W. A. Expanded natural product diversity revealed by analysis of lanthipeptide-like gene clusters in Actinobacteria. Appl. Environ. Microbiol. 2015, 81 (13), 4339−4350, doi: 10.1128/AEM.00635-15
[49] Zhang, Q.; Yu, Y.; Velasquez, J. E.; van der Donk, W. A. Evolutioń of lanthipeptide synthetases. Proc. Natl. Acad. Sci. U. S. A. 2012, 109 (45), 18361−18366, doi: 10.1073/pnas.1210393109
[50] Marsh, A.J., O'Sullivan, O., Ross, R.P. et al. In silico analysis highlights the frequency and diversity of type 1 lantibiotic gene clusters in genome sequenced bacteria. BMC Genomics 11, 679 (2010). https://doi.org/10.1186/1471-2164-11-679
[51] Repka L., Chekan J., Nair S., van der Donk W., Mechanistic Understanding of Lanthipeptide Biosynthetic Enzymes, Chemical Reviews, 2017, doi: 10.1021/acs.chemrev.6b00591
[52] Skaugen, M.; Andersen, E. L.; Christie, V. H.; Nes, I. F. Identification, characterization, and expression of a second, bicistronic, operon involved in the production of lactocin S in Lactobacillus sakei L45. Appl. Environ. Microbiol. 2002, 68 (2), 720−727, doi: 10.1128/AEM.68.2.720-727.2002
[53] Vinuesa P., Orchoa-Sanchez L., Complete Genome Sequencing of Stenotrophomonas acidaminiphila ZAC14D2_NAIMI4_2, a Multidrug-Resistant strain isolated from sediments of a polluted river in Mexico, uncovers new antibiotic resistance genes and a novel class-II Lasso peptide biosynthesis gene cluster, Genome Announc, 2015, 3(6):e01433-15, doi: 10.1128/genomeA.01433-15
[54] Stefanato F., Trippel C., Uszkoreit S., Ferrafiat L., Grenga L., Dickens R., Kelly N., Kingdon A., Ambrosetti L., Findlay K., Cheema J., Trick M., Chandra G., Tomalin G., Malone J., Truman A., Pan-genome analysis identifies intersecting roles for Pseudomonas specialized metabolites in potato pathogen inhibition, bioRxiv, 2019, doi: https://doi.org/10.1101/783258

[55] Ting C., Funk M., Halaby S., Zhang Z., Gonen T., van der Donk W., Use of a Scaffold Peptide in the Biosynthesis of Amino Acid derived Natural Products, Science, 2019, doi: 10.1126/science.aau6232
[56] Zdzislaw Z. E. Sikorski: Chemical and Functional Properties of Food Lipids. CRC Press, 2010, doi: 10.1201/9781420031997
[57] Kwon M., Steininger C., Cairns T., Wisecaver J., Lind A., Pohl C., Regner C., Rokas A., Meyer V., Beyond the biosynthetic gene cluster paradigm: genome-wide co-expression networks connect clustered and unclustered transcription factors to secondary metabolic pathway, bioRxiv, 2020, doi.: 10.1101/2020.04.15.040477
[58] Nouws S., Bogaerts B., Verhaegen B., Denayer S., Pierard D., Marchal K., Roosens N., Vanneste K., De Keersmaecker S., Impact of DNA extraction on whole genome sequencing analysis for characterization and relatedness of Shiga toxin-producing Escherichia coli isolates, Scientific Reports, 2020, 10(1):14649, doi: 10.1038/s41598-020-71207-3
[59] Instruction Manual, NEBNExt Ultra I and II, DNA Library Prep Kit for Illumina, Version 5.0, 2018
[60] http://core-genomics.blogspot.com/2014/01/nextseq-500s-new-chemistry-described.html, 15.08.2021
[61] https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html, 14.08.2021
[62] Hugerth L., Larsson J., Alneberg J., Lindh M., Legrand C., Pinhassi J., Andersson A., Metagenome-assembled genomes uncover a global brackish microbiome, Genome Biology, 2015, 16(279), doi: 10.1186/s13059-015-0834-7
[63] Alneberg J., Karlsson C., Divne A., Bergin C., Homa F., Lindh M., Hugerth L., Etterna T., Bertilsson S., Andersson A., Pinhassi J., genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single amplified genomes, Microbiome, 2018, 6(173), doi: 10.1186/s40168-018-0550-0
[64] Bolger, A. M., Lohse, M., & Usadel, B., Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170, 2014, 30(15):2114-2120, doi: 10.1093/bioinformatics/btu170
[65] Bankevich A., Nurk S., Antipov D., Gurevich A., Dvorkin M., Kulikov A. S., Lesin V., Nikolenko S., Pham S., Prjibelski A., Pyshkin A., Sirotkin A., Vyahhi N., Tesler G., Alekseyev M. A., Pevzner P. A. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology, 2012, 19(5):455-477, doi: 10.1089/cmb.2012.0021
[66] Li D., Liu C., Luo R., Sadakane K., Lam T., MEGAHIT: an ulta-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph, Bioinformatics, 2015, 31(10):1674-6, doi: 10.1093/bioinformatics/btv033
[67] Miller I., Rees E., Ross J., Miller I., Baxa J., Lopera J., Kerby R., Rey F., Kwan J., Autometa: automated extraction of microbial genomes from individual shotgun metagenomics, Nucleic Acids Research, 2019, 47(10):e57, doi: 10.1093/nar/gkz148
[68] Dick, G.J., Andersson, A.F., Baker, B.J. et al. Community-wide analysis of microbial genome sequence signatures. Genome Biol 10, R85 (2009). https://doi.org/10.1186/gb-2009-10-8-r85
[69] Teeling,H., Meyerdierks,A., Bauer,M., Amann,R. and Glockner, F.O., Application of tetranucleotide frequencies for the assignment of genomic fragments, 2004, Environ. Microbiol., 6(9): 938–947, DOI:10.1111/j.1462-2920.2004.00624.x
[70] Difrancesco P., Bonneau D., Hutchinson J., The implications of M3C2 projection diameter on 3D semi-automated rockfall extraction from sequential terrestrial laser scanning point clouds, Remote Sensing, 2020, 12(11):1885, DOI:10.3390/rs12111885
[71] Zhu X, Ghahramani Z. 2002. Learning from labeled and unlabeled data with label propagation. Technical report, CMU-CALD-02

[72] Kang D., Li F., Kirton E., Thomas A., Edgan R., An H., Wang Z., MetaBAT2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies, PeerJ, 2019, 7:e7359, doi: 10.7717/peerj.7359

[73] Parks D., Imelfort M., Skennerton C., Hugenholtz P., Tyson G., CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes, Genome Research, 2015, 25(7):1043-55, doi: 10.1101/gr.186072.114

[74] Gurevich A., Saveliev V., Vyahhi N., Tesler G., QUAST: quality assessment tool for genome assemblies, Bioinformatics, 2013, Vol. 29, 1072-1075, 29(8):1072-5, doi: 10.1093/bioinformatics/btt086

[75] Hyatt D., Chen G., LoCascio P., Land M., Larimer F., Hauser L., Prodigal: prokaryotic gene recognition and translation initiation site identification, BMC Bioinformatics, 2010, 11:119, doi: 10.1186/1471-2105-11-119

[76] W. C. Fuqua, S. C. Winans, E. P. Greenberg: Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators. Journal of bacteriology, 1994, 176(2):269-75, doi: 10.1128/jb.176.2.269-275.1994

[77] Papenfort, K., Bassler, B. Quorum sensing signal–response systems in Gram-negative bacteria. Nat Rev Microbiol 14, 576–588 (2016). https://doi.org/10.1038/nrmicro.2016.89

[78] Surette MG, Miller MB, Bassler BL. Quorum sensing in Escherichia coli, Salmonella typhimurium, and Vibrio harveyi: a new family of genes responsible for autoinducer production. Proc Natl Acad Sci U S A. 1999 Feb 16;96(4):1639-44. doi: 10.1073/pnas.96.4.1639

[79] Kim, C. S., Gatsios, A., Cuesta, S., Lam, Y. C., Wei, Z., Chen, H., Russell, R. M., Shine, E. E., Wang, R., Wyche, T. P., Piizzi, G., Flavell, R. A., Palm, N. W., Sperandio, V., & Crawford, J. M. (2020). Characterization of Autoinducer-3 Structure and Biosynthesis in E. coli. ACS Central Science, 6(2), 197-206. https://doi.org/10.1021/acscentsci.9b01076

[80] Vasquez JK, Blackwell HE. Simplified Autoinducing Peptide Mimetics with Single-Nanomolar Activity Against the Staphylococcus aureus AgrC Quorum Sensing Receptor. ACS Infect Dis. 2019 Apr 12;5(4):484-492. doi: 10.1021/acsinfecdis.9b00002

[81] Tan, C., Koh, K., Xie, C. et al. Community quorum sensing signalling and quenching: microbial granular biofilm assembly. npj Biofilms Microbiomes 1, 15006 (2015). https://doi.org/10.1038/npjbiofilms.2015.6

[82] Eric Block: Garlic and Other Alliums: The Lore and the Science. Royal Society of Chemistry, Cambridge 2010

[83] Scheller S, Goenrich M, Boecher R, Thauer RK, Jaun B. The key nickel enzyme of methanogenesis catalyses the anaerobic oxidation of methane. Nature. 2010 Jun 3;465(7298):606-8. doi: 10.1038/nature09015

[84] Partovi SE, Mus F, Gutknecht AE, Martinez HA, Tripet BP, Lange BM, DuBois JL, Peters JW. Coenzyme M biosynthesis in bacteria involves phosphate elimination by a functionally distinct member of the aspartase/fumarase superfamily. J Biol Chem. 2018 Apr 6;293(14):5236-5246. doi: 10.1074/jbc.RA117.001234

[85] Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, Suarez Duran HG, de Los Santos ELC, Kim HU, Nave M, Dickschat JS, Mitchell DA, Shelest E, Breitling R, Takano E, Lee SY, Weber T, Medema MH. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017 Jul 3;45(W1):W36-W41. doi: 10.1093/nar/gkx319

[86] Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W339-46. doi: 10.1093/nar/gkr466

[87] Glansdorff, N., Xu, Y. & Labedan, B. The Last Universal Common Ancestor: emergence, constitution and genetic legacy of an elusive forerunner. Biol Direct 3, 29 (2008). https://doi.org/10.1186/1745-6150-3-29
[88] Baum D., et.al., Reading a Phylogenetic tree: the meaning of monphyletic groups, Nature Education, 2008, 1(1):190
[89] Xiaoxia Liu, Jingxian Zhang, Feng Ni, Xu Dong, Bucong Han, Daxiong Han, Zhiliang Ji, Yufen Zhao: Genome-wide exploration of the origin and evolution of amino acids. BMC Evolutionary Biology. 2010, 10:77, doi: 10.1186/1471-2148-10-77
[90] Martini, M., Lee I., Bottner K., Zhao Y., Botti S., Bertaccini A., Harrison N., Carraro L., Marcone C., Khan A., Osler R., Ribosomal protein gene-based phylogeny for finer differentiation and classification of phytoplasmas, Int J Syst Evol Microbiol, 2007, 57(Pt 9):2037-51, DOI:10.1099/ijs.0.65013-0
[91] Yutin N., Puigbo P., Koonin E., Wolf Y., Phylogenomics of Prokaryotic Ribosomal Proteins, PLOS ONE, 2012, doi.org/10.1371/journal.pone.0036972
[92] Goris, J. et al. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. 2007, 57, 81–91, 57(Pt 1):81-91, doi: 10.1099/ijs.0.64483-0
[93] Kurtz, S., Phillippy, A., Delcher, A.L. et al. Versatile and open software for comparing large genomes. Genome Biol 5, R12 (2004). https://doi.org/10.1186/gb-2004-5-2-r12
[94] Richter M., Rossello-Mora R., Shifting the genomic gold standard for the prokaryotic species definition, PNAS, 2009, 106(45):19126-19131, https://doi.org/10.1073/pnas.0906412106
[95] Lee I, Ouk Kim Y, Park SC, Chun J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Microbiol. 2016 Feb;66(2):1100-1103. doi: 10.1099/ijsem.0.000760
[96] Yoon SH, Ha SM, Lim J, Kwon S, Chun J. A large-scale evaluation of algorithms to calculate average nucleotide identity. Antonie Van Leeuwenhoek. 2017 Oct;110(10):1281-1286. doi: 10.1007/s10482-017-0844-4
[97] Jain C., Rodriguez L., Phillippy A., Konstantinidis K., Aluru Srinivas, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries, Nature Communications, 2018, 9(1), DOI:10.1038/s41467-018-07641-9
[98] Emms, D.M., Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20, 238 (2019). https://doi.org/10.1186/s13059-019-1832-y
[99] Oscar Robinson, David Dylus, Christophe Dessimoz; Phylo.io : Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web . Mol Biol Evol 2016; 33 (8): 2163-2166. doi: 10.1093/molbev/msw080
[100] Yu G., Using ggtree to visualize data on tree-like structures, Current Protocols in Bioinformatics, 2020, https://doi.org/10.1002/cpbi.96
[101] Zhou Z., Lai J., Walsh C., Directed evolution of aryl carrier proteins in the enterobactin synthetase, PNAS, 2007, 104(28): 11621-11626, https://doi.org/10.1073/pnas.0705122104
[102] Motamedi H, Shafiee A, Cai SJ, Streicher SL, Arison BH, Miller RR. Characterization of methyltransferase and hydroxylase genes involved in the biosynthesis of the immunosuppressants FK506 and FK520. J Bacteriol. 1996 Sep;178(17):5243-8. doi: 10.1128/jb.178.17.5243-5248.1996
[103] Herzberg M, Kaye IK, Peti W, Wood TK. YdgG (TqsA) controls biofilm formation in Escherichia coli K-12 through autoinducer 2 transport. J Bacteriol. 2006 Jan;188(2):587-98. doi: 10.1128/JB.188.2.587-598.2006

[104] Wang L, Hashimoto Y, Tsao CY, Valdes JJ, Bentley WE. Cyclic AMP (cAMP) and cAMP receptor protein influence both synthesis and uptake of extracellular autoinducer 2 in Escherichia coli. J Bacteriol. 2005 Mar;187(6):2066-76. doi: 10.1128/JB.187.6.2066-2076.2005
[105] Frederick Verbeke, Severine De Craemer, Nathan Debunne, Yorick Janssens, Evelien Wynendaele, Christophe Van de Wiele, and Bart De Spiegeleer: Peptides as Quorum Sensing Molecules: Measurement Techniques and Obtained Levels In vitro and In vivo, Frontiers in Neuroscience, 2017, Band 11, S. 183, https://doi.org/10.3389/fnins.2017.00183
[106] Koonin E., Orthologs, Paralogs and Evolutionary Genomics, Annu. Rev. Genet. 2005, 39:309-38, doi: 10.1146/ annurev.genet.39.073003.114725
[107] Kanehisa M., Furumichi M., Tanabe M., Sato Y., Morishima K., KEGG: new perspectives on genomes, pathways, diseases and drugs, 2017, 45(Database issue):D353-D361, doi: 10.1093/nar/gkw1092
[108] Toronen P, Medlar A, Holm L, PANNZER2: A rapid functional annotation webserver. Nucl. Acids Res.46, 2018, W84-W88
[109] Louca S., Masel F., Doebeli M., Parfrey L., A census-based estimate of Earth's bacterial and archaeal diversity, 2019, doi.org/10.1371/journal.pbio.3000106
[110] Lin, YT., Lin, YF., Tsai, I.J. et al. Structure and Diversity of Soil Bacterial Communities in Offshore Islands. Sci Rep 9, 4689 (2019). https://doi.org/10.1038/s41598-019-41170-9
[111] Abraham, B.S., Caglayan, D., Carrillo, N.V. et al. Shotgun metagenomic analysis of microbial communities from the Loxahatchee nature preserve in the Florida Everglades. Environmental Microbiome 15, 2 (2020). https://doi.org/10.1186/s40793-019-0352-4
[112] Kielak A., Barreto C., Kowalchuk G., van Veen J., Kuramae E., The ecology of Acidobateria: Moving beyond genes and genomes, Front. Microbiol, 2016, https://doi.org/10.3389/fmicb.2016.00744
[113] Tessler, M., Neumann, J.S., Afshinnekoo, E. et al. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. Sci Rep 7, 6589 (2017). https://doi.org/10.1038/s41598-017-06665-3
[114] Amit V. Mangrola, Pravin Dudhagara, Prakash Koringa, C.G. Joshi, Rajesh K. Patel, Shotgun metagenomic sequencing based microbial diversity assessment of Lasundra hot spring, India, Genomics Data, Volume 4, 2015, Pages 73-75, https://doi.org/10.1016/j.gdata.2015.03.005.
[115] Chen R., Wong H., Kindler G., MacLeod F., Benaud N., Ferrari B., Burns B., Discovery of an Abundance of Biosynthetic Gene Clusters in Shark Bay Microbial Mats, Front. Microbiol., 2020, https://doi.org/10.3389/fmicb.2020.01950
[116] Crits-Christoph, A., Diamond, S., Butterfield, C.N. et al. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. Nature 558, 440–444 (2018). https://doi.org/10.1038/s41586-018-0207-y
[117] Santana-Pereira A., Sandoval-Powers M., Monsma S., Zhou J., Santos S., Mead D., Lites M., Discovery of Novel Biosynthetic Gene Cluster Diversity From a Soil Metagenomic Library, Front. Microbiol., 2020, https://doi.org/10.3389/fmicb.2020.585398
[118] Rudolf J., Yan X., Shen B., Genome Neighborhood Network Reveals Insights into Enediyne Biosynthesis and Facilitates Prediction and Prioritization for Discovery, J. Ind Microbiol Biotechnol., 2016, 43(0): 261-276, doi:10.1007/s10295-015-1671-0
[119] Bowers R., Kyrpides N., et.al., Minimum Information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea, Nat. Biotechnol., 2017, 35(8): 725-731, doi: 10.1038/nbt.3893
[120] Nowrotek M., Jalowiecki L., Harnisz M., Piaza G., Culturomics and metagenomics: in understanding of environmental resistome, Front. Environ, 2019, 13(3): 40, https://doi.org/10.1007/s11783-019-1121-8
[121] Kaur J., Sharma J., Orchid Root Associated Bacteria: Linchpins or Accessories?, Front. Plant. Sci., 2021, https://doi.org/10.3389/fpls.2021.661966

[122] Héctor Herrera, Inmaculada García-Romera, Meneses, C. , Pereira, G. , & César Arriagada. Orchid mycorrhizal interactions on the pacific side of the Andes from Chile. a review. Journal of Soil Science and Plant Nutrition, 2019, 19(1), 187-202

8 Code Availability

Codes, used for plotting in RStudio and data extraction

```
library(ggplot2)
library(ggtree)
library(pheatmap)

##### Completeness vs. Contamination #####
ggplot(checkm_results_1, aes(x = Completeness, y = Contamination)) +
  geom_point(alpha = 2, aes(color = Category)) +
  theme_bw() +
  ggtitle("Assessment of Metagnome Assembled Genome Quality") +
  labs(x = "Completeness [%]", y = "Contamination [%]") +
  scale_color_manual(values = c("#4E84C4", "black", "coral3"))

##### Completeness vs. Heterogeneity #####
ggplot(checkm_results_1, aes(x = Completeness, y = Heterogeneity)) +
  geom_point(alpha = 2, aes(color = Category)) +
  theme_bw() +
  ggtitle("Assessment of Metagnome Assembled Genome Heterogeneity") +
  labs(x = "Completeness [%]", y = "Heterogeneity [%]") +
  scale_color_manual(values = c("#4E84C4", "black", "coral3"))

##### Number of Contigs #####
ggplot(checkm_results_1, aes(x = reorder(MAG, -number_contigs), y = number_contigs)) +
  geom_bar(stat = "identity", aes(color = Category, fill = Category)) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("Number of Contigs") +
  labs(x = "MAG", y = "No. of Contigs") +
  scale_color_manual(values = c("#4E84C4", "black", "coral3")) +
  scale_fill_manual(values = c("#4E84C4", "black", "coral3"))

##### N50 #####
ggplot(checkm_results_1, aes(x = reorder(MAG, -n50), y = n50)) +
  geom_bar(stat = "identity", aes(color = Category, fill = Category)) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("N50") +
  labs(x = "MAG", y = "N50") +
  scale_color_manual(values = c("#4E84C4", "black", "coral3")) +
  scale_fill_manual(values = c("#4E84C4", "black", "coral3"))

##### L50 #####

ggplot(basic_stats_MAGs, aes(x = reorder(MAG, -L50), y = L50)) +
  geom_bar(stat = "identity", aes(color = Category, fill = Category)) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("L50") +
  labs(x = "MAG", y = "L50") +
  scale_color_manual(values = c("#4E84C4", "coral3", "black")) +
  scale_fill_manual(values = c("#4E84C4", "coral3", "black"))
```

```
##### GC content #####

ggplot(basic_stats_MAGs, aes(x = reorder(MAG, -GC_Content), y = GC_Content)) +
  geom_bar(stat = "identity", aes(color = Category, fill = Category)) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("GC Content") +
  labs(x = "MAG", y = "GC Content [%]") +
  scale_color_manual(values = c("#4E84C4", "coral3", "black")) +
  scale_fill_manual(values = c("#4E84C4", "coral3", "black"))

##### Average Coverage #####
ggplot(basic_stats_MAGs, aes(x = reorder(MAG, -Average_Coverage), y = Average_Coverage)) +
  geom_bar(stat = "identity", aes(color = Category, fill = Category)) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  ggtitle("Average Coverage") +
  labs(x = "MAG", y = "Average Coverage") +
  scale_color_manual(values = c("#4E84C4", "coral3", "black")) +
  scale_fill_manual(values = c("#4E84C4", "coral3", "black"))

##### unrooted tree #####
ggtree(tree, layout = "daylight", branch.length = 'none') +
  geom_tiplab(size = 3) +
  geom_hilight(node = 40, fill = "yellow") +
  xlim(-20, 20) +
  ylim(-20, 20)

##### ANI heatmap #####

Ani_matrix_class_1_2 <- read.delim("~/Desktop/Masterarbeit/ani_files/Ani_matrix_class_1_2.csv", row.names=1)
data1 <- as.matrix(Ani_matrix_class_1_2)
fontsize_row = 10 - nrow(data2) / 15
pheatmap(data2, main = "Average Nucleotide Identity Results", cluster_cols = F, cluster_rows = F, fontsize_row =
fontsize_row, border_color = NA)

##### Piechart for reference genome abundance #####

overview_species_abundance <- read.csv("~/Desktop/Masterarbeit/ani_files/overview_species_abundance.csv",
          row.names=1)
wght <- round(overview_species_abundance$Abundance/sum(overview_species_abundance$Abundance), digits = 2)
species <- rownames(overview_species_abundance)
ref.species <- as.data.frame(cbind(species, wght))
pie <- plot_ly(ref.species, labels = ~species, values = ~wght, type = 'pie', textposition = 'outside', textinfo = 'label+percent')
%>%
  layout(title = 'Reference Species Abundance for ANI-Analysis',
    xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
    yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))


##### Command to extract peptides based on their name #####
      ##### This command is an example for the biofilm specific peptide extraction #####
while read -r line; do awk -v pattern=$line -v RS=">" '$0 ~ pattern { printf(">%s", $0); }' final.contigs_amino_acid.fa; done <
./meta_peptides_in_pathways/meta_indices/meta_peptides_in_biofilm.txt > peptides_in_biofilm.txt
```

9 Supplementary Material

L50 values and GC content

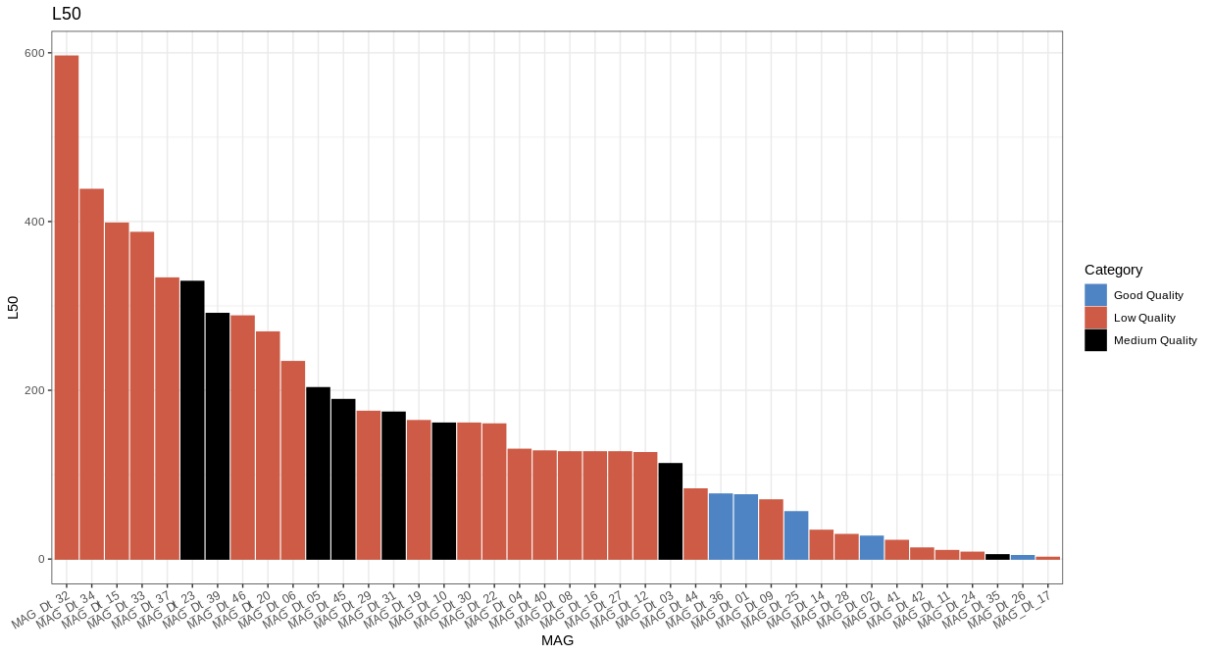


Figure 34. L50 values all MAGs, separated by quality. High quality MAGs have lower L50 values, which means that fewer contigs are needed to cover 50% of the genome.

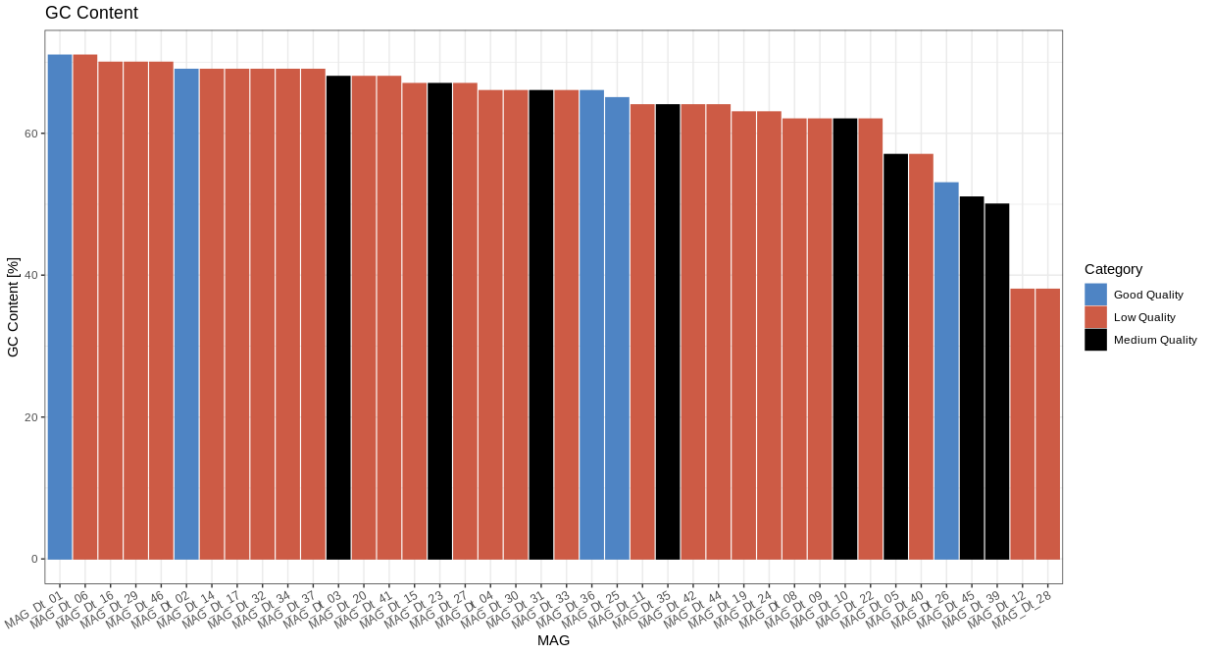


Figure 35. GC content in % for all MAGs, separated by quality. Low and medium quality MAGs could be difficult to interpret, due to the high incompleteness or contamination values.

Complete overview over the quality assessment of all MAGs.

Table 6.: Overview of completeness, contamination and heterogeneity level for all generated MAGs. Based on certain thresholds, MAGs were excluded or classified into low-, medium- or high-quality.

MAG	COMPLETENESS [%]	CONTAMINATION [%]	HETEROGENEITY [%]	CLASSIFICATION
MAG_Dt_01	80.24	4.58	27.91	High
MAG_Dt_02	94.87	11.76	60.00	High
MAG_Dt_03	65.94	3.99	77.27	Medium
MAG_Dt_04	14.73	0.00	0.00	Low
MAG_Dt_05	40.62	1.85	0.00	Medium
MAG_Dt_06	22.87	0.00	0.00	Low
MAG_Dt_07	0.00	0.00	0.00	Excluded
MAG_Dt_08	22.35	0.00	0.00	Low
MAG_Dt_09	16.67	0.00	0.00	Low
MAG_Dt_10	49.72	4.09	33.33	Medium
MAG_Dt_11	13.32	1.16	60.00	Low
MAG_Dt_12	13.23	0.00	0.00	Low
MAG_Dt_13	5.12	0.31	0.00	Excluded
MAG_Dt_14	21.57	2.20	44.44	Low
MAG_Dt_15	20.14	0.31	0.00	Low
MAG_Dt_16	20.81	1.75	100.00	Low
MAG_Dt_17	29.46	2.63	100.00	Low
MAG_Dt_18	4.71	0.00	0.00	Excluded
MAG_Dt_19	13.01	1.72	100.00	Low
MAG_Dt_20	12.07	1.72	0.00	Low
MAG_Dt_21	8.62	3.16	100.00	Excluded
MAG_Dt_22	13.08	1.63	33.33	Low
MAG_Dt_23	61.01	2.68	50.00	Medium
MAG_Dt_24	16.42	1.84	61.90	Low
MAG_Dt_25	90.35	22.87	91.67	High
MAG_Dt_26	92.03	2.53	50.00	High
MAG_Dt_27	26.65	0.58	0.00	Low
MAG_Dt_28	14.78	0.00	0.00	Low
MAG_Dt_29	25.30	0.00	0.00	Low
MAG_Dt_30	25.79	0.00	0.00	Low
MAG_Dt_31	34.21	13.88	66.67	Medium

MAG_Dt_32	15.50	0.00	0.00	Low
MAG_Dt_33	20.49	0.00	0.00	Low
MAG_Dt_34	25.79	0.00	0.00	Low
MAG_Dt_35	36.21	1.72	100.00	Medium
MAG_Dt_36	85.72	11.35	60.61	High
MAG_Dt_37	20.30	0.31	0.00	Low
MAG_Dt_38	1.72	0.00	0.00	Excluded
MAG_Dt_39	32.83	3.45	50.00	Medium
MAG_Dt_40	19.66	2.07	50.00	Low
MAG_Dt_41	14.04	0.00	0.00	Low
MAG_Dt_42	25.86	8.62	33.33	Low
MAG_Dt_43	100.00	966.68	2.48	Excluded
MAG_Dt_44	83.57	50.70	15.44	Low
MAG_Dt_45	38.49	2.49	30.00	Medium
MAG_Dt_46	27.18	33.33	32.14	Low
MAG_Dt_47	96.39	137.01	24.44	Excluded

Complete overview over the BGCs of all MAGs.

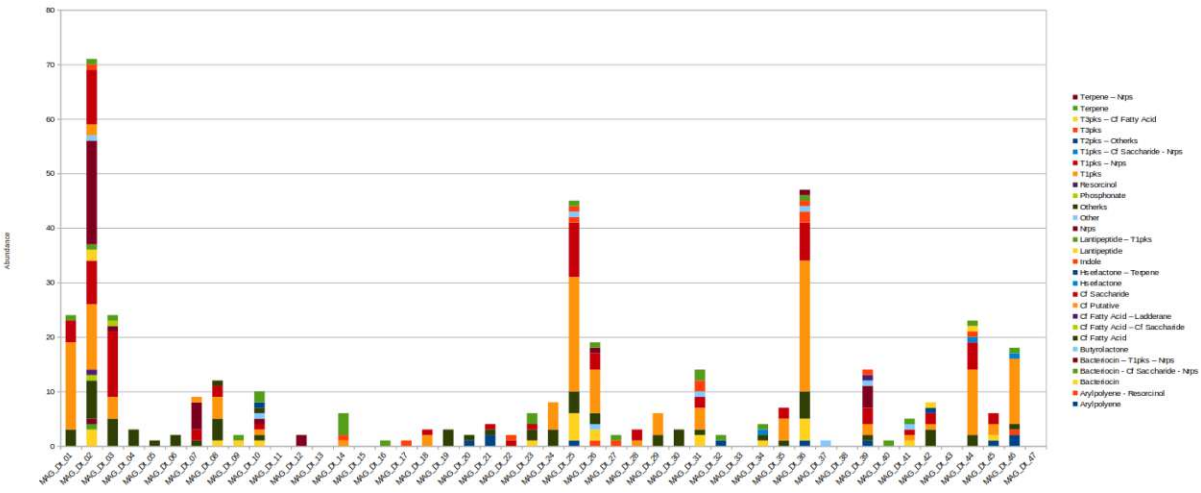


Figure 36. Overview of the discovered BGCs, sorted by type and MAG.

Putative Alkanesulfone Biosynthesis as a reaction to oxidative stress.

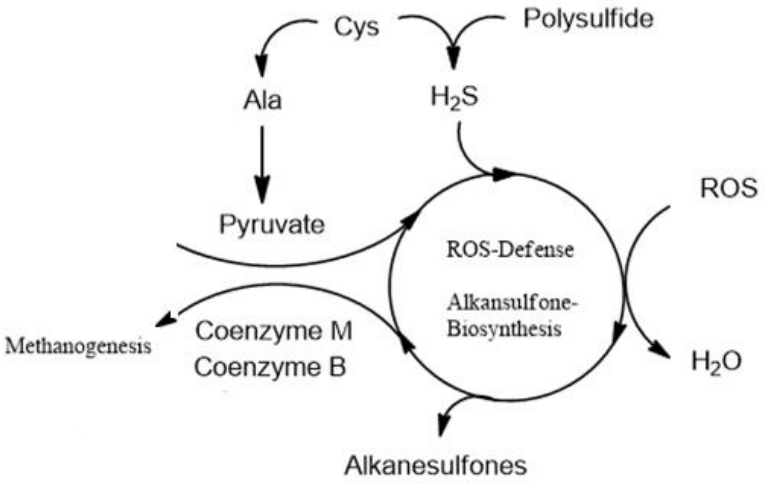
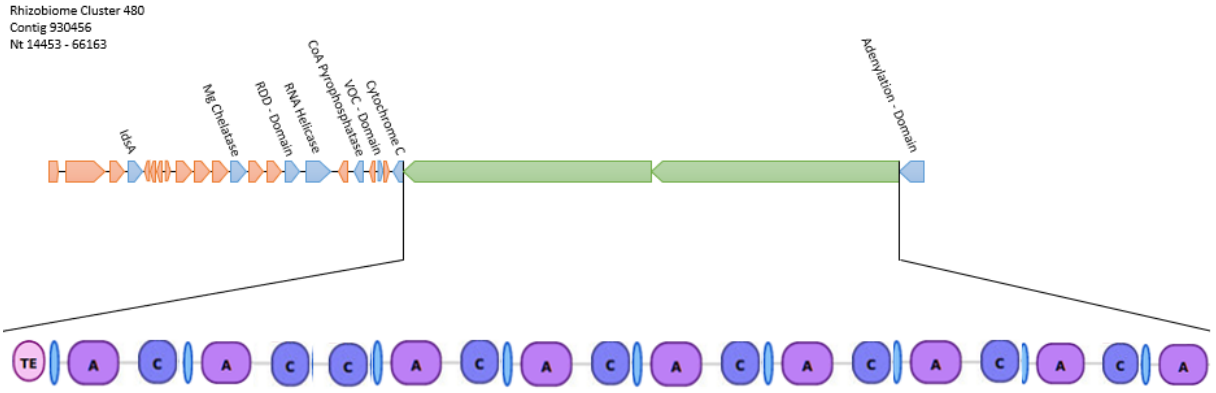


Figure 37. Hypothetical connection between oxidative stress response and alkanesulfon biosynthesis, located within the 4Fe-4S containing peroxidoredoxin.

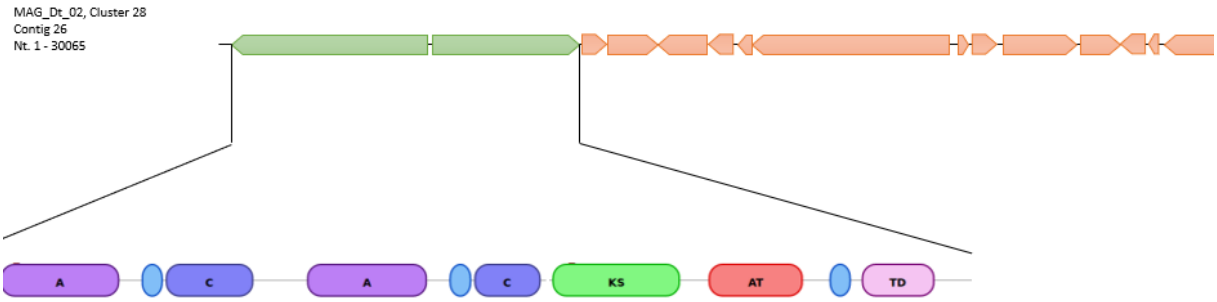
NRPS - Cluster



Legend: Rhizobiome, Cluster 480, Contig 930456, Nt. 14453 – 66163; Core genes 29 (nt. 34453-48267) and 30 (nt. 48376-64341)
Domain: A = Adenylation, C = Condensation, TE = Thioesterase

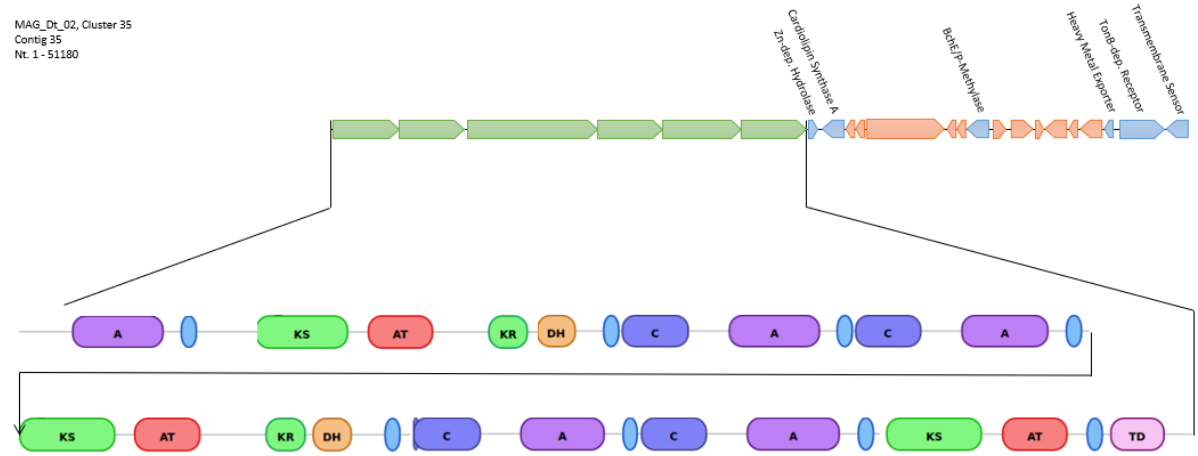
Figure 38. NRPS-BGC containing complete modules, but an extra condensation domain in one module. The genes are divided into core (green), annotated (blue) and unannotated (red) genes.

Hybride NRPS-PKS Cluster



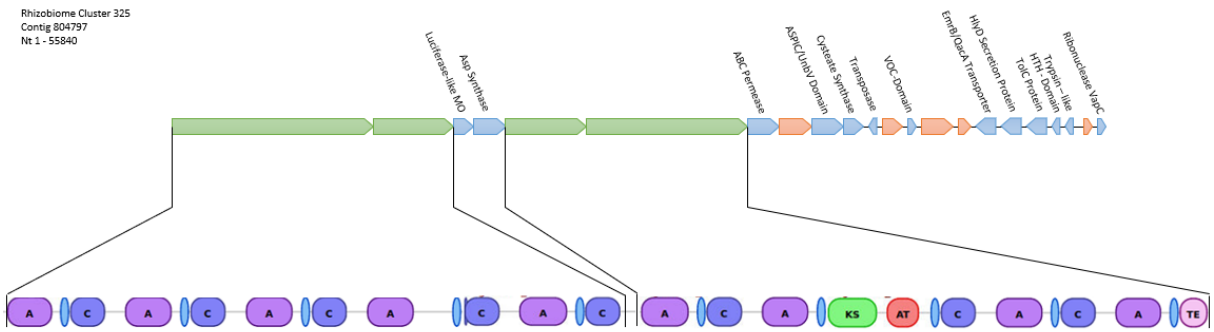
Legend: MAG_Dt_02, Cluster 28, Contig 26, Nt. 1 – 30065; Core genes 1 (nt. 1-5625) and 2 (nt. 5752-10065)
Domain: A = Adenylation, C = Condensation, KS = Ketosynthase, AT = Acyltransferase, TD = Terminal Reductase

Figure 39. Hybrid NRPS-PKS BGC, containing complete modules. The genes are divided into core (green), annotated (blue) and unannotated (red) genes.



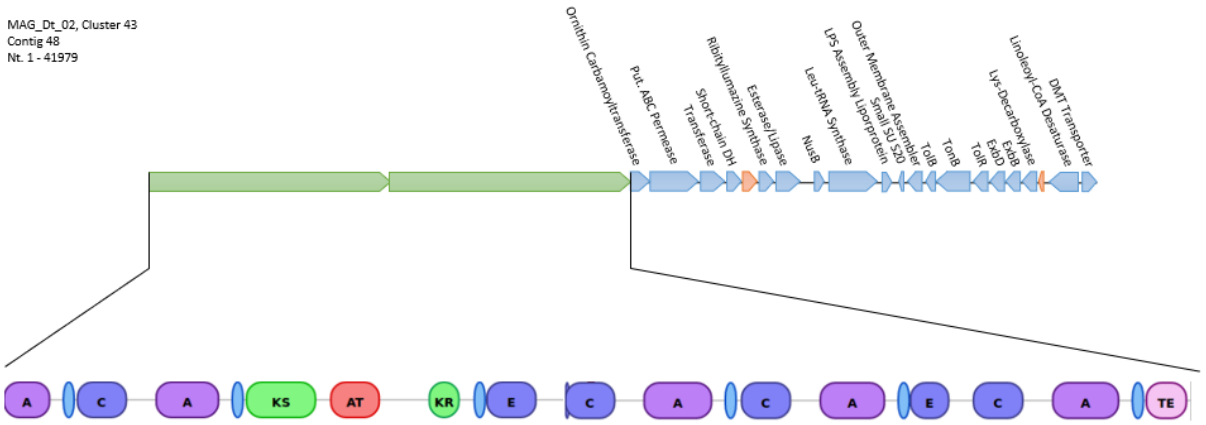
Legend: MAG_Dt_02, Cluster 35, Contig 35, Nt. 1 – 51180; Core genes 1 (nt. 1-4005), 2 (nt. 4002-7931), 3 (nt. 7919-15397), 4 (nt. 15394-20700), 5 (nt. 20697-27020) and 6 (nt. 27017-31180)
Domain: A = Adenylation, C = Condensation, KS = Ketosynthase, AT = Acyltransferase, E = Epimerization, KR = Ketoreductase, DH = Dehydratase, TD = Terminal Reductase

Figure 40. Hybrid NRPS-PKS BGC, containing complete modules. The genes are divided into core (green), annotated (blue) and unannotated (red) genes.



Legend: Rhizobiome, Cluster 325, Contig 804797, Nt. 1 – 55840; Core genes 1 (nt. 1-13065), 2 (nt. 13062-17660), 3 (nt. 17663-18799), 4 (nt. 18801-20669), 5 (nt. 20683-25728) and 6 (nt. 25731-35840)
Domain: A = Adenylation, C = Condensation, KS = Ketosynthase, AT = Acyltransferase, TE = Thioesterase

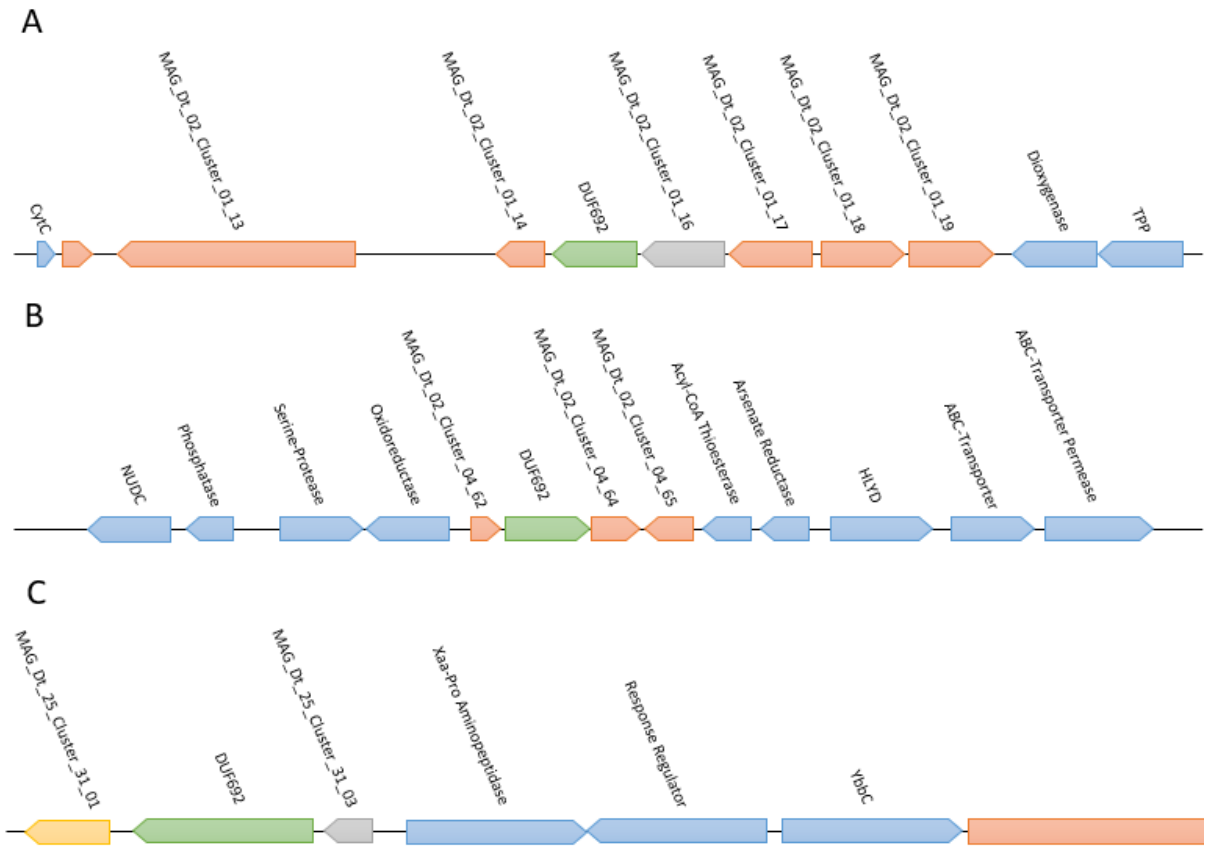
Figure 41. Hybrid NRPS-PKS BGC, containing complete modules. The genes are divided into core (green), annotated (blue) and unannotated (red) genes.



Legend: MAG_Dt_02, Cluster 43, Contig 48, Nt. 1 – 41979; Core genes 1 (nt. 3-10382) and 2 (nt. 10379-21979)
Domain: A = Adenylation, C = Condensation, KS = Ketosynthase, AT = Acyltransferase, KR = Ketoreductase, TE = Thioesterase, E = Epimerization

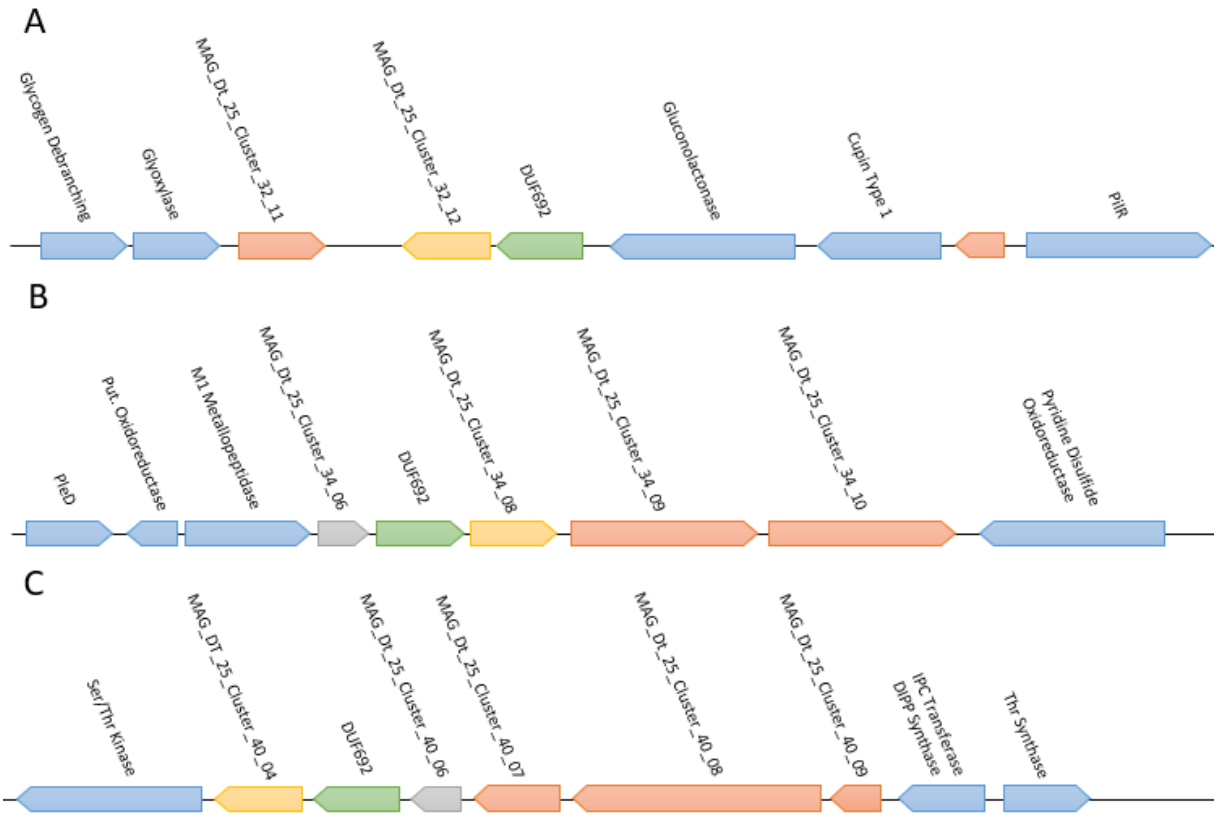
Figure 42. Hybrid NRPS-PKS BGC, containing complete modules. The genes are divided into core (green), annotated (blue) and unannotated (red) genes.

Annotated Bacteriocin BGCs, defined by the DUF692-domain containing protein.



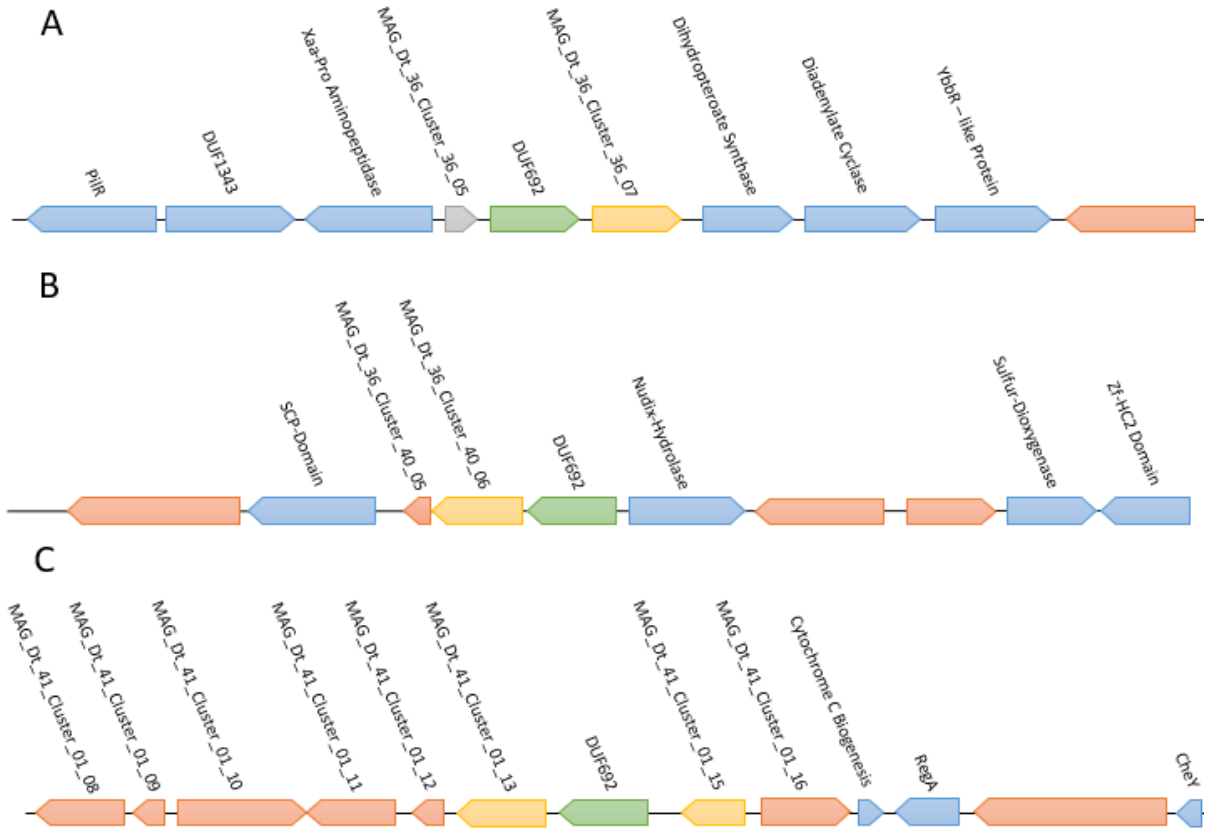
Legend: A) MAG_Dt_02, Cluster 01, Contig 01, Nt. 12158 – 23006
B) MAG_Dt_02, Cluster 04, Contig 03, Nt. 61056 – 71976
C) MAG_Dt_25, Cluster 31, Contig 92, Nt. 1 – 6529

Figure 43. BGCs containing the DUF692-protein as core protein. The genes are divided into core (green), annotated (blue) and unannotated (red) genes. Genes in yellow or grey were discovered to cluster together after the multiple sequence alignment. Grey genes could potentially be silver efflux pumps due to the fact that they cluster together with an annotated silver efflux pump.



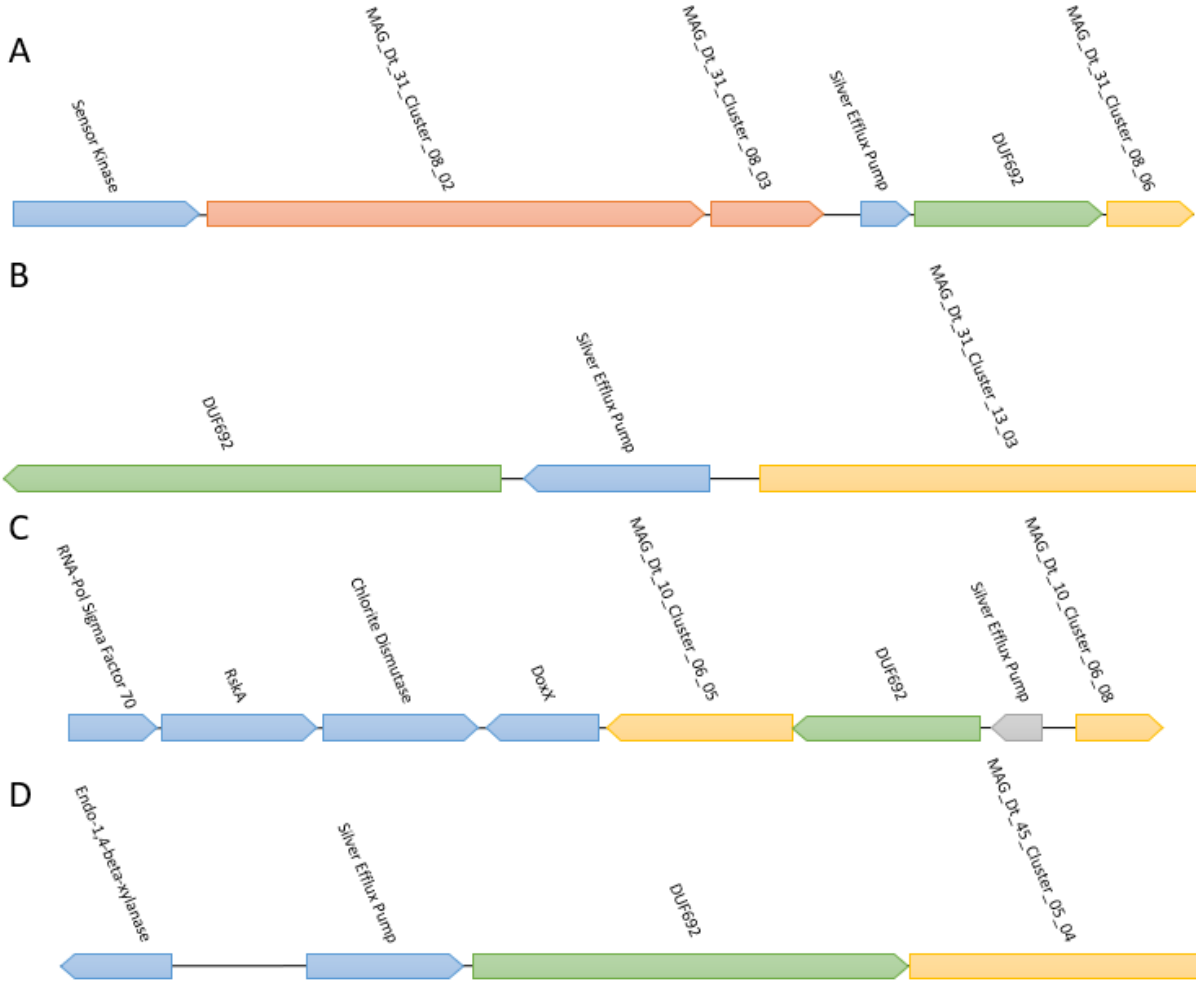
Legend: A) MAG_Dt_25, Cluster 32, Contig 99, Nt. 10475 – 21302
B) MAG_Dt_25, Cluster 34, Contig 104, Nt. 747 – 11577
C) MAG_Dt_25, Cluster 40, Contig 143, Nt. 2127 – 13029

Figure 44. BGCs containing the DUF692-protein as core protein. The genes are divided into core (green), annotated (blue) and unannotated (red) genes. Genes in yellow or grey were discovered to cluster together after the multiple sequence alignment. Grey genes could potentially be silver efflux pumps due to the fact that they cluster together with an annotated silver efflux pump.



Legend: A) MAG_Dt_36, Cluster 36, Contig 167, Nt. 357 – 11316
B) MAG_Dt_36, Cluster 40, Contig 196, Nt. 2739 – 13554
C) MAG_Dt_41, Cluster 01, Contig 29, Nt. 6679 – 17566

Figure 45. BGCs containing the DUF692-protein as core protein. The genes are divided into core (green), annotated (blue) and unannotated (red) genes. Genes in yellow or grey were discovered to cluster together after the multiple sequence alignment. Grey genes could potentially be silver efflux pumps due to the fact that they cluster together with an annotated silver efflux pump.



Legend: A) MAG_Dt_31, Cluster 08, Contig 129, Nt. 1 – 5016
B) MAG_Dt_31, Cluster 13, Contig 462, Nt. 1 – 2133
C) MAG_Dt_10, Cluster 06, Contig 155, Nt. 1 – 5202
D) MAG_Dt_45, Cluster 05, Contig 9571, Nt. 1 – 2489

Figure 46. BGCs containing the DUF692-protein as core protein. The genes are divided into core (green), annotated (blue) and unannotated (red) genes. Genes in yellow or grey were discovered to cluster together after the multiple sequence alignment. Grey genes could potentially be silver efflux pumps due to the fact that they cluster together with an annotated silver efflux pump.

Table 7.: Overview of the renamed genes for the multiple sequence alignment and clustering of the according to the similarity scores.

NUMBER	GENE IN BGC	ANNOTATION	CLUSTER (HEATMAP)	LOCATION (NT.)
X1	MAG_Dt_10_Cluster_06_8	-	Cluster_3	4794-5201
X2	MAG_Dt_26_Cluster_10_80	-	Cluster_3	95992-96729
X3	MAG_Dt_25_Cluster_40_4	-	Cluster_3	6295-7143
X4	MAG_Dt_36_Cluster_30_7	-	Cluster_3	4952-5737
X5	MAG_Dt_41_Cluster_01_15	-	Cluster_3	12950-13540
X6	MAG_Dt_31_Cluster_13_3	-	Cluster_3	1236-2132
X7	MAG_Dt_10_Cluster_06_5	-	Cluster_3	2363-3250
X8	MAG_Dt_25_Cluster_31_1	-	Cluster_3	1-684
X9	MAG_Dt_36_Cluster_36_7	-	Cluster_3	6313-7125
X10	MAG_Dt_26_Cluster_15_100	-	Cluster_3	104566-105342
X11	MAG_Dt_45_Cluster_05_4	-	Cluster_3	1834-2487
X12	MAG_Dt_36_Cluster_40_6	-	Cluster_3	6843-7742
X13	MAG_Dt_41_Cluster_01_13	-	Cluster_3	10795-11682
X14	MAG_Dt_25_Cluster_34_8	-	Cluster_3	6574-7503
X15	MAG_Dt_25_Cluster_32_12	-	Cluster_3	14630-15478
X16	MAG_Dt_31_Cluster_08_6	-	Cluster_3	4650-5015
X17	MAG_Dt_36_Cluster_40_5	-	-	6435-6797
X18	MAG_Dt_41_Cluster_01_8	-	-	6878-7720
X19	MAG_Dt_41_Cluster_01_12	-	-	10331-10732
X20	MAG_Dt_26_Cluster_10_82	-	-	97592-97867
X21	MAG_Dt_02_Cluster_01_18	-	-	19751-20761
X22	MAG_Dt_02_Cluster_04_65	-	-	67554-68015
X23	MAG_Dt_31_Cluster_08_2	-	-	709-2904
X24	MAG_Dt_02_Cluster_01_16	-	Cluster_2	17978-18919
X25	MAG_Dt_25_Cluster_40_6	-	Cluster_2	7127-8029
X26	MAG_Dt_36_Cluster_30_4	-	Cluster_2	3544-3900
X27	MAG_Dt_26_Cluster_15_98	-	Cluster_2	103425-103682
X28	MAG_Dt_10_Cluster_06_7	Silver Efflux Pump	Cluster_2	4228-4518
X29	MAG_Dt_25_Cluster_34_6	-	Cluster_2	5378-5674
X30	MAG_Dt_25_Cluster_31_3	-	Cluster_2	1531-1806
X31	MAG_Dt_36_Cluster_36_5	-	-	4829-5146
X32	MAG_Dt_26_Cluster_15_96	-	-	101983-102441
X33	MAG_Dt_26_Cluster_15_97	-	-	102455-103033

X34	MAG_Dt_25_Cluster_40_8	-	-	9369-10718
X35	MAG_Dt_36_Cluster_30_2	-	-	1094-2599
X36	MAG_Dt_25_Cluster_32_11	-	-	13587-14213
X37	MAG_Dt_25_Cluster_40_9	-	-	10697-11083
X38	MAG_Dt_41_Cluster_01_9	-	-	7859-8215
X39	MAG_Dt_31_Cluster_08_3	-	-	2901-3362
X40	MAG_Dt_02_Cluster_04_64	-	-	66973-67575
X41	MAG_Dt_02_Cluster_01_14	-	-	16689-17171
X42	MAG_Dt_02_Cluster_01_13	-	-	13149-14873
X43	MAG_Dt_25_Cluster_40_7	-	-	8596-9372
X44	MAG_Dt_36_Cluster_30_3	-	-	2596-3393
X45	MAG_Dt_41_Cluster_01_16	-	-	13705-14727
X46	MAG_Dt_41_Cluster_01_11	-	-	9507-10334
X47	MAG_Dt_25_Cluster_34_9	-	-	7605-9659
X48	MAG_Dt_25_Cluster_34_10	-	-	9668-11497
X49	MAG_Dt_02_Cluster_01_19	-	-	20778-21731
X50	MAG_Dt_02_Cluster_01_17	-	-	18916-19647
X51	MAG_Dt_41_Cluster_01_10	-	-	8512-9582
X52	MAG_Dt_36_Cluster_30_5	-	-	3907-4110
X53	MAG_Dt_02_Cluster_04_62	-	-	65812-66012
X54	MAG_Dt_02_Cluster_04_63	DUF692-Protein	Cluster_1	66056-66976
X55	MAG_Dt_02_Cluster_01_15	DUF692-Protein	Cluster_1	17158-18006
X56	MAG_Dt_25_Cluster_40_5	DUF692-Protein	Cluster_1	7127-8029
X57	MAG_Dt_36_Cluster_30_6	DUF692-Protein	Cluster_1	4123-4968
X58	MAG_Dt_26_Cluster_10_81	DUF692-Protein	Cluster_1	96713-97567
X59	MAG_Dt_36_Cluster_40_7	DUF692-Protein	Cluster_1	4952-5737
X60	MAG_Dt_41_Cluster_01_14	DUF692-Protein	Cluster_1	11679-12566
X61	MAG_Dt_25_Cluster_32_13	DUF692-Protein	Cluster_1	15475-16302
X62	MAG_Dt_31_Cluster_08_5	DUF692-Protein	Cluster_1	3794-4660
X63	MAG_Dt_31_Cluster_13_1	DUF692-Protein	Cluster_1	2-844
X64	MAG_Dt_26_Cluster_15_99	DUF692-Protein	Cluster_1	103709-104566
X65	MAG_Dt_45_Cluster_05_3	DUF692-Protein	Cluster_1	989-1834
X66	MAG_Dt_25_Cluster_31_2	DUF692-Protein	Cluster_1	681-1529
X67	MAG_Dt_36_Cluster_36_6	DUF692-Protein	Cluster_1	7739-8554
X68	MAG_Dt_25_Cluster_34_7	DUF692-Protein	Cluster_1	5747-6577
X69	MAG_Dt_10_Cluster_06_6	DUF692-Protein	Cluster_1	3274-4140